

If it ain't broke, don't fix it (much)

Modest structural reforms could help the US National Institutes of Health to maintain its independence — and the public's confidence.

Employees at the National Institutes of Health (NIH) in Bethesda, Maryland, have been chafing lately under what many regard as unwarranted micromanagement by officials at the Department of Health and Human Services. And some members of Congress are casting an unusually critical eye over the agency: in June, one committee began investigating awards given to Richard Klausner in 1999, when he directed the National Cancer Institute. It broadened its investigation after press reports that an NIH microbiologist had received a fat salary since 1995 for doing almost no work.

It is in this context that the National Academies' Institute of Medicine (IOM) has just issued a report on the NIH's organizational structure. The IOM committee came up with a range of solutions to address the agency's various issues. The scientific community, if it wants to see the NIH retain its annual budget of almost \$30 billion, should work closely with disease advocates and members of Congress to see that some of these measures are implemented.

The IOM grappled with three main problems. The first is an ever-increasing number of institutes with more specialized scientific portfolios. Some who have observed the NIH at close quarters say that consolidating some of the institutes would allow the agency to fund and conduct more coordinated and cutting-edge projects.

However, much of the agency's generous budget increases have resulted from lobby groups pushing for funding for their favourite institutes. Getting rid of the super-specialized structure would be politically challenging, as these lobby groups are well supported in Congress, where any reform measure must pass muster. The IOM rejects any general effort at consolidation as more trouble than it's worth. This isn't particularly brave, although it may well be prudent.

But the IOM report does recommend two specific mergers: between the human genome and general medical-sciences institutes,

and between the institutes of alcohol and drug abuse. Both proposals have considerable merit, although it is far from clear whether Congress will ratify even this relatively minor consolidation.

The problems posed by the proliferation of NIH institutes over the years are compounded by the weakness of the director's office. The IOM recommends giving the director more money for cross-NIH research projects and for a special-projects programme, initially worth \$100 million, that would be loosely modelled on the Defense Advanced Research Projects Agency (DARPA), to support high-risk research. These proposals would stimulate work in emerging fields and in areas that cross institute boundaries, as well as giving the director more clout in persuading institutes to cooperate.

Finally, the report tackles the issue of political influence over the NIH's activities. Many of the IOM's recommendations aim to tweak the balance between politicians, scientists and NIH officials. The committee's suggestion that all institute directors be appointed by the NIH director — not by the health secretary — would merely formalize existing practice. Its proposal to reform the external advisory system is laudable, as the current patchwork of arrangements doesn't always expose senior NIH officials to enough viewpoints.

More contentious is the idea that institute directors have their term in office limited to five years, with possible renewal for only five more after review. While meeting concerns that some institutes lack sufficiently dynamic leadership, this might also remove experienced individuals, exposing the institutes to more political interference.

The IOM's ideas are hardly revolutionary: it issued similar findings in 1984, but they were ignored. The scientific community and Congress, which is likely to consider new authorization legislation for the NIH next year, should instigate some intelligent reforms, which could help forestall future management problems at the agency. ■

Welcome to the Anthropocene

The sizzling, soaraway summer mustn't get in the way of European climate scientists' objectivity.

Meteorologists, atmosphere researchers and climate modellers throughout Europe have a summer job this year. Reporters and television crews are queuing up for interviews, demanding explanations about whether the current heatwave means that man-made climate change has finally arrived.

The heat produces strange fancies in some quarters. On 8 August, for example, Germany's best-selling newspaper, the *Bild*, splashed this on its front page: "Heat researchers alarmed: Equator dramatically shifted." That kind of thing doesn't do much for public understanding of science. But in these overheated days, it is essential that scientists and journalists find the right balance in telling people what they should expect, or even fear, in tomorrow's weather.

Some may be tempted to use the heatwave to ram home the fact that climate change is probably with us, but researchers must not overshoot the mark in their public statements. Unusually warm summers, after all, occur from time to time anyway.

But there's nothing wrong with scientists taking the opportunity to

remind the public and policy-makers that climate change is real. Global mean temperatures have increased by about 1 °C over the past hundred years, and will probably keep rising. Evidence is growing that this trend will lead to changes in atmospheric circulation that correlate with increased occurrence of extreme weather events, such as storms, droughts, floods and heatwaves (see *Nature* **421**, 805; 2003).

Humans have had little experience of climate change during our brief, recorded history. Throughout the Holocene — the past 11,000 years or so — global mean temperatures have been pretty stable. But during glacial and inter-glacial times, for example, people had to adapt to substantial climate fluctuations.

As we enter a period that climate researchers have dubbed the 'Anthropocene', we must develop strategies to mitigate the impact of climate change on health, safety and prosperity. Those who can afford it can take out insurance. Perhaps the current heatwave in wealthy Western Europe will build public support for a global fund to help the billions of people in poor countries who don't have that choice. ■





In from the cold
Thawing permafrost
helps to gauge
climate change
p712



Ground control
Optical astronomy
poses Australian
challenge
p714



Killer cure
Biologists seek to get
mammal numbers
back on track
p715



Shot down
Siberian grizzly
study falls foul of
poachers' aim
p716

Chinese fusion method promises fresh route to human stem cells

Carina Dennis

Biologists in China have reprogrammed human cells by fusing them with rabbit eggs emptied of their genetic material. And they have extracted stem cells, which have the potential to form a wide array of different cell types, from the resulting embryos.

The researchers, led by Huizhen Sheng of Shanghai Second Medical University, think that these 'derived' stem cells could provide scientists in the field with an alternative to stem-cell lines extracted from human embryos. But some researchers who have seen the work point out that the derived cells don't seem to have the same ability as human embryonic stem cells to grow indefinitely in culture.

The work will be published online this week in *Cell Research*, a peer-reviewed journal supported by the Chinese Academy of Sciences. The paper will appear in print later this month (Y. Chen *et al. Cell Res.* **13**, 251–263; 2003).

Sheng's work has already created a buzz after rumours of it circulated in the scientific community and were reported in *The Wall Street Journal* in March 2002 (see *Nature* **419**, 334–336; 2002). The publication is likely to reignite debate over the ethics of cross-species reprogramming. But cell biologists say that having the data available for public discussion will help researchers and regulators to decide what kinds of cross-species work should be pursued.

Reprogramming adult cells to assume an embryonic state could offer a way to grow new cells and tissues to replace those lost to ageing and disease. And using an individual's own cells, in a process often called therapeutic cloning, may avoid problems of the immune system rejecting the cell therapy.

To reprogramme an adult cell, it can either be fused with or have its nucleus injected into an egg that has had its own nucleus removed. The reconstituted cell is then tricked into dividing as if it were an embryo, with all memory of its previous life as a liver, skin or kidney cell erased. After 5–7 days, embryonic stem cells — which can seemingly grow indefinitely — can be extracted from the growing ball of cells. By



changing the growth conditions, these cells can then be coaxed to develop into many different cell types.

Until now, scientists have only been able to generate animal stem-cell lines from the reprogrammed nuclei of cells. "This is the first paper to convincingly show that you can get human reprogramming," says Robin Lovell-Badge, a cell biologist at the National Institute for Medical Research in London.

Sheng claims to have successfully reprogrammed cells from the foreskin tissues of males aged 5, 42 and 52 years and from the facial skin of a 60-year old woman. "It just goes to show that age doesn't matter," says Lovell-Badge.

But Sheng has had a tough time convincing some experts. "I first submitted the paper more than two years ago," she says. It is understood that the paper was considered and rejected by other journals before this week's publication in China.

"I was frustrated that it took so long to get the paper published," Sheng says, "and it still may take a while for people to accept the work. But the scientific community has the right to question the details of the work and we have a responsibility to respond to them."

Doug Melton a cell biologist at Harvard University, for one, has concerns about the



A team led by Huizhen Sheng (left) has devised a technique for reprogramming adult human cells by fusing them with empty rabbit eggs (above).

nature of the derived embryonic stem-cell lines. "I'm convinced that the cells do have the capacity to differentiate into different cell types, but it's unclear how long the cells can grow in culture," he says.

"It would be very surprising if the cell lines were stable," Melton adds, noting that many interspecies hybrids are unstable, because of incompatibilities between the nucleus and mitochondria (the energy-producing compartments of cells) from different species.

And Rudolf Jaenisch, of the Whitehead Institute in Cambridge, Massachusetts, is not convinced that the derived cells meet the usual criteria for embryonic stem cells. "An important criterion is indefinite growth," he says. "This is not shown."

But reservations aside, Melton is pleased that the work has finally seen the light of day. "I'm glad to see it published as it will encourage others to try it," he says.

At this stage, Sheng has no plans to use stem cells created by her method to treat humans. "It is a research tool. But there is the possibility that if it were proved to be safe enough for clinical use, it could provide a solution to human egg shortages for reprogramming in the future," says Sheng.

While the scientific community debates her work, Sheng is keen to test the embryonic stem cells she has generated in an animal model. "It will be important to see whether they will be tolerated by the immune system and whether they can correct an animal model of human disease," she says.

www.cell-research.com

Alpine thaw breaks ice over permafrost's role

Quirin Schiermeier, Munich

As Europe swelters in one of its hottest summers since records began, the heat is making itself felt in the high Alps, where thawing ice is destabilizing landscapes that are normally frozen.

Researchers who study permafrost say that the patterns of disruption being seen in the Alps right now could soon be replicated on a larger scale, with global warming eroding the quarter of the Earth's land surface that is permanently frozen.

In the Alps, most ground above an altitude of 2,500 metres — including about 6% of Switzerland — remains frozen throughout the year, with only a surface layer less than 5 metres thick melting in the summer and re-freezing in winter. Above 3,000 metres, temperatures rarely exceed 0 °C for any length of time during the summer. In the past few weeks, however, ice has been thawing out steadily at altitudes of up to 4,600 metres.

Slopes containing ice-filled cracks and discontinuities have become unstable, rendering the area unsafe for climbers. On 15 July, for example, nearly 100 people were taken off the Matterhorn — Switzerland's best-known peak — by helicopter, after a rockslide had blocked their descent. At least 50 climbers have been killed this summer by Alpine rockfalls.

"Unfortunately, it is only incidents such as that at the Matterhorn that seem to raise awareness of the importance of permafrost research," says Daniel Vonder Mühll,



Digging deep: Swiss permafrost probes could provide clues to long-term climate change.

a geophysicist at the University of Basel in Switzerland.

Marcia Phillips, a geographer at the Swiss Federal Institute for Snow and Avalanche Research in Davos, investigates long-term permafrost changes in Switzerland, their impact on slope failure, and the stability of structures designed to hold back avalanches in Alpine valleys.

The institute is involved in PERMOS, a scheme to monitor permafrost temperature and slope stability at 26 drilling sites throughout the Swiss Alps. "There is a large

variability, due to differences in topography. But by and large the threat of slope failure is increasing," says Phillips.

Permafrost thawing is not restricted to the Alps. In northern Europe and Russia, for example, thawing permafrost is a growing challenge to the construction and operation of roads, bridges and oil and gas pipelines.

Climatologists see permafrost as a sensitive lead indicator of climate change. Researchers in the European Union-funded Permafrost And Climate in Europe (PACE) project have found evidence of temperature increases of between 0.5 and 2.0 °C over the past 60–80 years in permafrost soils in European mountain regions, from the Sierra Nevada in Spain to the arctic archipelago of Spitsbergen (C. Harris (ed.) *Permafrost and Periglacial Processes* 12 (1); 2001).

"There is clear evidence for accelerated thawing in the upper permafrost layer in many regions," says Lorenz King, a geographer at Justus Liebig University in Giessen, Germany. If heavy snow in the winter insulates the ground from low air temperatures, permafrost can thaw even more deeply in summer, says Phillips.

Scientists have also suggested that permafrost melting can release vast amounts of carbon into the atmosphere, as the organic material that it contains is broken down by bacteria (see *Nature* 409, 751; 2001). This raises the spectre of a vicious circle in which these increasing quantities of greenhouse gases drive climate change, melting more permafrost and accelerating global warming. ■

Flowers' sick trick gives botanists double trouble

John Whitfield, London

Sickly plants have fooled botanists into incorporating the symptoms of their ailments into species descriptions, researchers say. The discovery raises the prospect that some plants currently classified as separate species are really just unhealthy samples of other groups.

"We need to continually validate the accuracy of species descriptions," says botanist Michael Hood of the University of Virginia, Charlottesville, co-author of a paper published last week that points out the error.

Hood and his colleague Janis Antonovics use herbarium specimens to map the distribution of plant diseases. As part of their study, they examined specimens of the genera *Silene* and *Lychnis* (which belong to the same family as carnations) in the herbarium of the National Museum of Natural History in Paris. The plants were collected in China about a century ago.

Of the 300 species examined, specimens of 25 were found to be infected with a fungus called anther smut. This causes the plant's anthers — the male sex organs — to become filled with dark purple spores instead of yellow pollen. None of the specimens had been noted as unusual by their collectors (M. E. Hood and J. Antonovics *Proc. R. Soc. Lond. B* (suppl.) doi: 10.1098/rsbl.2003.0063; 2003).

The diseased plants included the 'type specimens' for four species. A type specimen is the one on which a species' scientific description is based. *Silene cardiopetala*, for example, is distinguished from *Silene tatarinowii* by its dark anthers. Hood suspects that the plants may be one and the same. A flora of China published in 2001 repeats this description of *S. cardiopetala*.

"It's a cautionary tale," says Nicholas Turland, a specialist in Chinese plants at the Missouri Botanical Garden, St Louis. "It

shows that when you describe a new species, you should collect lots of material, and not base it on a few specimens." Rare mutations could also produce errors, he points out.

Hood and Antonovics have so far failed to find similar mistakes in other plant groups. "How often this occurs is difficult to say," Hood says. "Either taxonomists are very good at distinguishing diseased specimens, or they are still hidden out there."

The disease discovery throws fresh light on a major problem area in taxonomy, botanists say. Bogus species names often arise when a researcher redescribes something that has already been named, or when different variants of a species — such as flower colour — are given separate names. The International Plant Names Index contains more than one million entries, but researchers think that there are really only between 200,000 and 400,000 species of flowering plant. ■

Plant-disease experts blighted by red tape

Jonathan Knight, San Francisco

Concerns over homeland security have put US plant-disease scientists in the spotlight, and they're enjoying more money and greater visibility as a result. But the attendant bureaucracy and costly security measures already have some of them asking if all the attention is a blessing or a curse.

"It really has been a double-edged sword," says Jacqueline Fletcher, president of the American Phytopathological Society, which took up the issue at its annual meeting in Charlotte, North Carolina, this week.

Plant pathology has always been a poor relation of human health science, commanding only a tiny fraction of the government funding that biomedical research attracts. But since 11 September 2001, the people who study the blights, rusts, bacteria and viruses that attack plants have found themselves at the centre of a drive to protect America's food supply from potential terrorist acts. "There has been a greater understanding of plant pathology as something important to investigate," says Fletcher.

But with that visibility has come a growing concern about who has access to the pathogens under investigation. The Agricultural Bioterrorism Protection Act of 2002, passed last December, includes a list that was drawn up by the Animal and Plant Health Inspection Service (APHIS), of ten plant pathogens for which researchers must now undergo security checks if they are to continue studying them.

Although researchers recognize that stricter regulation is necessary, in certain cases it is doing more harm than good, says Caitilyn Allen, a plant pathologist at the University of Wisconsin, Madison. The



Green dilemma: could pathogens such as potato late blight prove to be dangerous in the wrong hands?

bacterium she studies, *Ralstonia solanacearum*, is on the APHIS list.

Although a major pest of potatoes in developing countries, *Ralstonia* is not a problem in North America, and Allen doubts it ever could be. It does poorly in northern climates, and has caused only minor losses in parts of Europe where it is endemic. It has also entered the United States several times via imported geraniums without consequence. "This is not the Ebola virus of the plant world," she points out.

Yet her university has spent \$50,000 adding locks and cameras to her lab to comply with the new security rules. The lab must also keep a continuous log of the amount of *Ralstonia* on hand, no easy task for an organism that doubles overnight. Although Allen has taken the trouble to comply, at least three of her colleagues at other universities have

not, she says, choosing instead to destroy their *Ralstonia* collections.

Jan Leach at Kansas State University in Manhattan, Kansas, says she nearly gave up working on the rice pathogen she studies, *Xanthomonas oryzae* pv. *oryzicola*, when it was listed by APHIS. She says existing that monitoring and inspection measures were sufficient. Not only are the nearest rice paddy-fields hundreds of miles away, but the disease causes only minor problems anyway.

One concern with the new requirements, says Leach, is that plant-disease labs will be less ready to respond when a new disease does arrive. At least two major crop diseases — the potato late blight (*Phytophthora infestans*) and the wheat infection karnal bunt (*Tilletia indica*) — have entered the United States since 1985, each prompting new research efforts.

Megan Thomas, a spokeswoman for APHIS, says the agency is planning to consult researchers and to re-evaluate the list annually, but she does not know when possible changes will be proposed.

Other researchers are starting to question whether there are elements of their research that they should not publish. In 1989, Martin Dickman of the University of Nebraska in Lincoln found that adding a single gene to a fungus that infects papaya through breaks in the skin gave it the ability to penetrate the fruit's skin on its own. Such work helps pathologists to understand what makes microbes virulent, but it could also in theory help terrorists to engineer nastier bugs. Asked if he would publish the finding today, he says: "I would have to think about it a lot harder."

Plant pathologists are not alone in facing these challenges, says Janet Shoemaker, director of public and scientific affairs for the American Society for Microbiology. After the anthrax attacks of 2001 there was a big push for new restrictions. "What we don't know yet is whether it has gone too far," she says.

Gold-mine conversion eases closer

Geoff Brumfiel

Officials in the state of South Dakota say that they have resolved a key liability issue that is blocking the conversion of the abandoned Homestake gold-mine into a deep underground laboratory.

Scientists want to convert the gold-mine, which is located near the city of Lead, into a state-of-the-art underground laboratory for studying neutrinos, rare physical processes, and geomicrobiology. But the mine's owner, Barrick Gold Corporation of Toronto, Canada, has refused to hand over the mine unless the company is indemnified against environmental liability (see *Nature* 415, 105; 2002).

Last week, the South Dakota state government opened offices for the

Homestake Laboratory Conversion Project, which it hopes will eventually assume responsibility for the mine. The authority, among other things, plans to take out an insurance policy on the 127-year-old mine, alleviating the concerns of its current owners. "This authority will enable us to move ahead," says Richard Gowen, who heads the project from its main office in Rapid City.

Barrick spokesman Vince Borg says the company believes that the new proposal is a "likely route" for transferring the mine, and that the two sides are drawing up guidelines for donating the mine. But he adds that the handover will not take place unless the National Science Foundation approves funding for the project.

Astronomers count the cost of thinking big

Carina Dennis, Sydney

Optical astronomy is now dominated by the word 'large', which is posing a sizeable problem for many countries that want to remain at its cutting edge.

Increasingly, ground-based projects are centred on big, expensive endeavours as exemplified by several proposals currently being circulated for an extremely large telescope (ELT). But medium-sized countries such as Australia face an uphill struggle to achieve meaningful participation in these large, internationally funded projects.

Australian astronomers are keenly aware of the dilemma, and a task force set up in May by the National Committee for Astronomy (NCA) is seeking a way for the country to establish a prominent role in an ELT project.

"We need to be part of a large telescope programme to survive," says Ken Freeman, an astronomer based at the Australian National University (ANU) in Canberra.

"There is an urgent need for Australia to decide what it wants to do, how it does it and with whom," says Matthew Colless, who is chairing the task force. The panel aims to help prepare a case that could convince a sceptical government to fork out upwards of A\$100 million (US\$65 million) for participation in an ELT project.

The NCA, part of the Australian Academy of Science, is preparing its decadal review of astronomy priorities, which it will deliver to the government within two years. The task force's report is expected to help set some of those priorities.

"We'll need to make some tough decisions over the next year about the science we want to be doing over the next decade," says Chris Tinney, acting director of the Anglo-Australian Observatory (AAO), headquartered in Sydney.

The Southern hemisphere has a built-in advantage for astronomers — the centre of the Milky Way can best be seen from that half of the planet. But Australian observatories cannot compete with the higher altitude sites in the Andes, where future optical telescopes are likely to be built.

Australian astronomers want their government to spend at least A\$100 million to participate in such a telescope project, which will cost of the order of US\$1 billion in total. They hope to secure one-off funding for the construction, but operations could eat into the country's astronomy research budget of about A\$40 million a year.

The task force is holding informal discussions with colleagues in Europe, the United States and Canada to explore options for the ELT project. "Japan would also be a very attractive overseas partner," says Colless.

But Australia is not about to abandon its existing facilities to pursue international



The Anglo-Australian Observatory, which recorded these star trails, has had to redefine its priorities.

projects. Local telescopes are being upgraded to carry out surveys of the entire southern sky, including the Radial Velocity Experiment and 6dF Galaxy Survey. "We depend on Australian observatories to do these essential surveys," says Richard McMahon, an astronomer at the University of Cambridge, UK.

Nevertheless, the AAO — one of Australia's main observatories — was hit hard two years ago by a British decision to withdraw financial support after 2006 to concentrate on the European Southern Observatory in Chile (see *Nature* 414, 678; 2001). At the moment, Britain provides half of the AAO's annual A\$8-million budget.

To compensate for its impending shortfall, the AAO is honing its skills in instrument building in anticipation of supplying equipment to the European Southern Observatory, among others. "The United Kingdom is far from pulling out of the AAO," Tinney contends. After 2006, he says, Australia and Britain will remain partners, but the observatory's focus "will move away from the support of current telescopes, and into instrument building for larger telescopes".

Australia is already recovering some of the costs it incurred when it became a member of Gemini — the US-led international observatory based in Hawaii and Chile, which it joined in 1998 — through instrumentation contracts with other Gemini participants. "It's clear we can be a small-percentage player but gain huge returns,"

says Penny Sackett, director of the ANU's Mount Stromlo Observatory, which is building two instruments for Gemini.

The AAO, meanwhile, has built equipment to position optical fibres for the European Southern Observatory's Very Large Telescope, and is currently making a similar tool for the Subaru telescope in Hawaii. The Subaru machine will be the first of its kind, says Tinney. The fibres, which carry the telescope's signals, are usually put in place one by one using robot arms, but the new system encases them inside 'spines' and positions them in parallel. "Instead of taking an hour to position the fibres, we can do it in minutes," says Tinney.

Such shifts in emphasis may be helping Australia's observatories to survive, but an ELT will be harder to finance. Tinney says that redirecting funds from existing astronomy projects or even shutting entire facilities wouldn't cover the costs of joining an ELT project. "We need substantial new funds," he says.

But getting such funding for an optical telescope won't be easy — particularly when radioastronomers in Australia are also hoping to host two major facilities, a low-frequency array and a square-kilometre array. The astronomy community insists it is united behind the home-based radio telescopes and the optical one overseas. "It's not a question of either/or — it's a question of which foot forward first," says Rachel Webster, of the University of Melbourne and chair of the NCA. ■

Bush accused of power abuse over science

Erika Check, Washington

A prominent Congressman has accused the Bush administration of interfering with US science to such an extent that it is threatening public trust in both science and government.

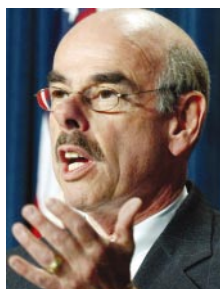
Henry Waxman (Democrat, California) says that the administration has blocked the dissemination of scientific information, interfered with research results or sought undue influence in the composition of advisory panels, in its handling of a range of issues including AIDS and climate change.

On 7 August, Waxman's staff released a 33-page report, *Politics and Science in the Bush Administration*, detailing claims that President George Bush and his officials were improperly involving themselves in a variety of aspects of the scientific process.

The report covers allegations of interference at agencies including the National Institutes of Health (NIH), the Centers for Disease Control and Prevention, the Food and Drug Administration (FDA) and the Environmental Protection Agency. The report alleges that most of this has catered to a conservative or pro-business agenda. For instance, it recounts an attempt by the Department of Health and Human Services to appoint a doctor who opposed abortion to a key FDA committee on reproductive health.

"This report shows there is a pattern here that cannot be ignored," Waxman says. "What we're talking about here is unprecedented."

Kathryn Harrington, a spokeswoman for



Henry Waxman: Bush's efforts to influence science are 'unprecedented'.

"We rely on people we elect not to use their power in ways that inappropriately distort knowledge," says Sheila Jasanoff, author of *The Fifth Branch: Science Advisers as Policymakers* (Harvard Univ. Press, Cambridge, Massachusetts, 1990) and a historian at Harvard University's Kennedy School of Government. "The strongest critique of the Bush administration is that this idea of delegation is being violated."

Critics say that the Bush administration's cavalier treatment of science has emboldened others to try to influence the scientific process. Last month, for example, Congress narrowly defeated a proposal by a Republican representative that would have blocked the NIH from funding five grants dealing with aspects of HIV, sexual health and behaviour, and wild animal populations. "The

the White House Office of Science and Technology Policy, says that the report is unfair. "This administration does indeed look at the facts and reviews the best available science to make decisions based on what is best for the American people," Harrington says.



interference is now down to the grant level, and that should really be a cause of concern to the scientific community," says Gregg Gonsalves of New-York-based Gay Men's Health Crisis.

Scientists who have spoken out welcomed the report. They claim that the Bush administration's attempts to influence science put it in danger of becoming irrelevant on major issues related to science.

"After a while, people just won't believe the administration because, again and again, the policies are flying in the face of the facts," says Phil Coyle, assistant defence secretary until 2001, and now a consultant at the Washington-based Center for Defense Information. ■

♦ www.politicsandscience.org

Biologists call for tracking as mammal numbers dive

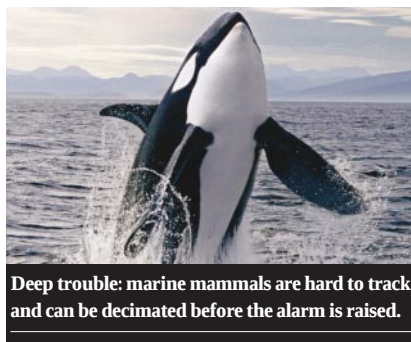
Rex Dalton, Portland

An agenda for marine-mammal research is being developed in the United States to try to save a range of often-elusive species from possible extinction.

More than 50 experts from around the world met in Portland, Oregon, on 4–7 August to consider priorities for research on dwindling populations of mammals such as whales, porpoises, sea lions, seals, otters, manatees and polar bears.

The US Marine Mammal Commission sponsored the meeting and will make recommendations later to Congress, which is considering changes to the Marine Mammal Protection Act. The US Navy and some commercial interests want to reduce protection, whereas marine biologists want to strengthen it.

Speakers at the Portland meeting backed development of a comprehensive monitoring and systematic sampling network to identify health issues and aid remedial plans. The creation of reservations in US waters —



Deep trouble: marine mammals are hard to track and can be decimated before the alarm is raised.

where research could be done and regulatory regimes tried out — was also proposed.

"I am encouraged by the research ideas at the workshop," says John Hildebrand, who studies mammal acoustics at the Scripps Institution of Oceanography in La Jolla, California. "But I'm discouraged by the magnitude of the problems."

For more than 20 years, marine mammals have been protected by US federal law. But several species have been hit by factors such

as loss of habitat, pollution, ocean noise and deaths inflicted by watercraft. The problems are worst for species, such as whales, that traverse unprotected international waters.

Jay Barlow, a marine biologist at the Southwest Fisheries Science Center in La Jolla, told the meeting it was so difficult to chart marine-mammal populations that 90% of a species could be lost before researchers realized it was in crisis.

Daniel Goodman, a zoological statistician at Montana State University in Bozeman, said that tracking needed new funding, not money from existing programmes.

There was little confidence that the Bush administration or the Republican-controlled Congress would provide the money. But scientists are optimistic that new research initiatives could be created.

On 19 August, a congressional subcommittee will hold a hearing in San Diego, where populations of seals and sea lions are so thriving that some residents and fishers are expected to call for a cull. ■

Head of Australian stem-cell centre quits to return to science

Sydney Developmental biologist Alan Trounson has stepped down as chief executive of Australia's National Stem Cell Centre to concentrate on his research.

Trounson will retain some management involvement as vice-chairman of the centre, which is to be built at Monash University's Clayton campus, but says that he intends to spend most of his time doing science. "He wants to put his energies back into scientific matters," says Hugh Niall, who will act as chief executive until a permanent appointment is made.

Trounson, a leader of embryonic stem-cell research in Australia who campaigned hard for the A\$43.5-million (US\$30-million) centre's establishment, came under heavy fire last year for allegedly misleading parliamentarians. In testimony before them, he showed a video of a paralysed rat recovering after the injection of what he said were embryonic stem cells — but which were in fact embryonic germ cells (see *Nature* **419**, 4; 2002).

Funding for the centre was then suspended while an independent inquiry reviewed the process that selected it, but the project was given a final go-ahead last December.



Alan Trounson: will retain some involvement in the stem-cell centre he helped to set up.

Open-access trial next step in biology journal's life cycle

London Oxford University Press (OUP) has announced that it will pilot an 'open-access' publishing model for one of its leading journals, *Nucleic Acids Research*.

Next January's annual database issue of the journal — a widely used reference guide of review articles about databases used in the life sciences — will be freely available online to non-subscribers. The costs of the open-access experiment will be covered in part by the authors, who will be charged

Grizzly end for environmentalists' bear project

London Poachers have ended the work of two Canadian conservationists who were studying grizzly bears (right) in eastern Siberia.

Charlie Russell and his wife Maureen Enns returned to their cabin on the Kamchatka Peninsula this spring to find a bear's gall bladder nailed to their kitchen wall. There were no other signs of the bears the couple had been studying, and other evidence suggested that human poachers had killed them during the winter.

The couple, who have been working in the area for seven years, say they are unsure who was responsible for the killing. But they concede

that local hunters were angry over their efforts to set up a wildlife sanctuary to protect the grizzlies.



T. BRAKEFIELD/CORBIS

about \$500 for each paper published.

If the experiment proves successful, the publisher says it will gradually move to an extended open-access model over the next four or five years, where subscription revenues are replaced by author charges.

"We feel that research should be globally accessible," says Martin Richardson, director of OUP's journals division. "But for open access to become a global reality, funding practices will have to change, and grant-givers will have to provide scientists with money to cover their publication costs."

Traditional values for transatlantic venture

Washington The University of Oxford and the Scripps Research Institute at La Jolla, California, are setting up a joint graduate programme in biochemistry — the first venture of its kind for both institutions.

Richard Lerner, president of Scripps, says that the collaboration grew out of talks between himself and Raymond Dwek, head of biochemistry at Oxford. "We were discussing how to bring traditional values to the high-tech, equipment-intensive world that chemistry and biology have found themselves in, and we decided each side could bring something special to this programme," Lerner says.

The joint programme — which will be funded out of a US\$100-million donation from the family of Samuel Skaggs, a drugstore-millionaire and philanthropist — is the latest in a spate of transatlantic alliances between leading research institutions.

"I think this will lay the foundations for significant interactions, and may even lead to joint appointments and research projects," says Dwek.

US neuroscience institute fills directorship at last

Washington A two-and-a-half-year vacancy at one of the largest institutes in the US National Institutes of Health (NIH) was filled last week, as Story Landis was

appointed to head the National Institute of Neurological Disorders and Stroke (NINDS).

Landis came to NINDS in 1995 to run its intramural science programme. She previously chaired the department of neuroscience at Case Western Reserve University School of Medicine, in Cleveland, Ohio, where she researched the development of the nervous system.

"The most important challenge for this institute in the next couple of years is going to be to achieve the right balance between basic science and clinical trials," Landis says, adding that she hopes to strengthen links with other NIH institutes that work on issues involving the nervous system.

Landis will take up her new post at the \$1.5-billion institute on 1 September. But there are still three directorships open at the NIH. They include that of the National Institute of Environmental Health Sciences, whose director, Kenneth Olden, announced late last month that he would leave when a replacement is identified.

Arizona State University plans clear-out of faculty

San Diego In a sweeping move designed to open up new positions, Arizona State University (ASU) is offering early retirement to a quarter of its faculty members.

About 570 faculty and professional staff at the university's campus at Tempe near Phoenix, out of a total of 2,100, are eligible for the early retirement programme. The programme is the brainchild of ASU's president Michael Crow, who arrived last year from Columbia University in New York.

"I believe ASU must significantly strengthen and expand its research activities, both to raise our national profile and to secure additional sources of funding," Crow said in a statement announcing the programme.

Staff have until December to consider the offer, after which ASU administrators will begin working out future hiring plans.

Do you want to work here?

The architects behind a new generation of laboratories believe their designs can stimulate good science. Laura Bonetta finds out how, and looks at research that may one day help to test their claims.



In 1992, the late virologist Jonas Salk was asked to accept an award from the American Institute of Architects for the institute he founded in La Jolla, California. In his speech, Salk told how he had visited an abbey in Assisi, Italy, while working on a polio vaccine early in his career. Although he did not go into details, he claimed that something about the architectural experience helped him to overcome difficulties in his work. In 1954, his vaccine underwent successful national trials in the United States, and eventually helped to cut the number of polio cases in the country by tens of thousands.

Are such stories one-offs, or can the physical environment really influence scientific creativity? Laboratory designers are increasingly banking on the hope that it can. Some are trying to maximize the chance encounters between scientists that spark new directions in research. Others are designing buildings intended to operate like real neighbourhoods, where workers experience a sense of community within a large building. "A lot more thought is going into the architecture of scientific buildings," says Robert McGhee, who has been resident architect with the Howard Hughes Medical Institute (HHMI) for more than 30 years.

Perhaps more importantly, at least in the long term, architects are also thinking about how they can test the claims that they make for the buildings. What does it mean to say that a building enhances creativity? Can this assertion be evaluated? A collaboration between architects and neuroscientists, established this spring, is aiming to answer such questions. Work is at an early stage, but those involved hope eventually to understand how architecture — both good and bad — affects our bodies and minds.

The current attention to good lab architecture has its roots in the Salk Institute. After his success with the polio vaccine, Salk decided to create an independent centre for biological research, and believed that architecture could play an important part in making it a success. During the late 1950s and 1960s, he worked with Louis Kahn, one of the twentieth century's most influential architects, to design an environment that was both welcoming and inspiring. "The Salk was the starting point for using the building to support the scientific community rather than just provide space," says McGhee.

Before then, labs were treated in much the same way as classroom buildings, with a corridor and often windowless rooms down

each side. The Salk Institute was a departure because Kahn had a high regard for natural light and coupled it with an understanding of the functional requirements of a lab. The outer four walls of the laboratory floors are glass, creating an open and airy environment. Levels containing equipment such as centrifuges and glassware are located in floors between the labs, freeing up space for research and giving scientists room to adapt their working environment to new experiments.

Pride of place

Equally importantly, the Salk Institute is architecturally adventurous, and is regarded by many as a stunning building. It is a source of pride for those who work there, an effect that university officials are now taking into account when commissioning buildings that they hope will attract the best scientists (see *Naturejobs*, page 858). "Everyone who works at the Salk has an intuition that it is a special place," says Fred Gage, a neuroscientist at the institute. "The design is externally aesthetic. It is very hard to pin down, but it has an effect."

Many science buildings designed since have mimicked these design features. Take the HHMI's Janelia Farm facility, for example. The new multidisciplinary research institute



TOP, SALK INSTITUTE; ABOVE, FRANK GEHRY

Breaking ground: the Salk Institute (top), the California NanoSystems Institute (top right), the Max Planck Institute's Dresden centre (right) and the Ray and Maria Stata Center (left), with its informal 'streets' (above), are part of an innovative new crop of research buildings.

is being built into a hill on the banks of the Potomac River near Leesburg, Virginia, and is set to open in 2005. Labs at Janelia Farm are positioned along one side of a glass corridor that runs the length of the undulating building. The glazed walls and ceiling will establish a visual connection to the landscape, while also providing access to the open terraces and offices that alternate along the length of the plan opposite the labs. "It is totally different from being in a box," says McGhee, who has worked on the design with New York-based firm Rafael Viñoly Architects.

In recent years, architects have begun to pay attention to a newer concept — the need for interactive spaces in which scientists can meet and talk to each other. The trend has been fuelled partly by the rise of fields such as nanotechnology, which require scientists from different disciplines to work together. Researchers who have traditionally been scattered across a university campus are now being brought together under a single roof. "People who would not normally speak are put in an environment that stimulates interactions," says James Broach, a former acting director of the Lewis-Sigler Institute for Integrative Genomics at Princeton University, New Jersey. "The hope is that sparks will fly."

Most architects agree that vertical distance is more of a barrier than horizontal distance when it comes to researcher interaction. Atria and open staircases, which create sight lines between floors, are one solution. In the Max Planck Institute of Molecular Cell Biology and Genetics in Dresden, Germany, built in 2001, the atrium occupies the full height and width of the four-story building. Heikkinen-Komonen Architects, based in Helsinki, Finland, designed the building so that suites of labs and offices radiate from the atrium and are connected to its central staircase by concrete bridges, which the scientists call *Ponte Vecchios*, after the famous bridge in the Italian city of Florence. A café and dining area, overlooking a rear garden, occupy the ground floor. "The mission was to have as many chance encounters as possible," says Kai Simons, the institute's executive director.

A similar approach has been taken by Rafael Viñoly's firm in its work on the California NanoSystems Institute at the University of California, Los Angeles, which is due to be completed at the end of next year. The architects faced a particular challenge at this site, as the three levels of laboratory space will be built around an existing car park. To ensure that the parking areas do not impair

movement between the labs, the design incorporates a series of zigzagging ramps that connect across and between floors.

But creating a building that increases interaction between researchers is not enough on its own. Scientists also need space and tools to help them bounce ideas off each other. With this in mind, architects design special areas, equipped with whiteboards, that should give researchers comfortable discussion space away from the bench. Deciding where to put them, however, can be tricky. "They work very well if they are in the right places," says McGhee. "But if you put them at the end of buildings, for example, they are used to get away rather than interact."

Traffic management

At the Stowers Institute for Medical Research in Kansas City, designed by Peckham Guyton Albers & Viets, also based in Kansas City, and completed in 2000, the common rooms are in areas of high traffic — for example, where shared equipment such as imaging hardware is housed. "Labs often have the equipment across the hall," says the institute's scientific director, Robert Krumlauf. "At the Stowers you have to go to shared areas." In a similar vein, the Genentech Hall building, which opened in January 2003 at the new Mission Bay campus of the University of California, San Francisco (UCSF), uses staircase landings to foster discussions. "If you are waiting at the elevator or see someone as you are using the stairs, you can flag them down and sit to have a conversation," says Keith Yamamoto, whose genetics group moved into the building when it opened.

Food and drink also help to promote discussion, as long as the environment is right. In 1994, the Imperial Cancer Research Fund in London, now part of Cancer Research UK, added a pub to its canteen area — but staff chose to continue using other local venues, because the pub offered no added convenience or incentive for socializing, and it was removed. Other similar initiatives may prove to be more successful. At the Lewis-Sigler Institute, which opened last autumn, a critical focal point is the coffee shop, which has wireless Internet connection, in the building's atrium. "People gather there," says Broach. "It is an opportunity to see someone you would not normally see."

But big, multidisciplinary buildings also come with problems. In a centre that employs hundreds of scientists, it is easy to feel lost. To avoid this, some architects are modelling lab layouts on real cities, with 'street' areas for interaction and 'neighbourhoods' where researchers can retreat into smaller groups. "The trend is to think about the design of a building in the way you think of cities," says Steven Wiesenthal, campus architect for UCSF.

The scientists involved in designing

LEFT, RAFAEL VIÑOLY ARCHITECTS; BELOW LEFT, HEIKKINEN-KOMONEN

Genentech Hall, for example, borrowed the neighbourhood concept that had been perfected in labs at the UCSF's original Parnassus campus. "The research towers there were 16 stories high," says Yamamoto of the old campus. "So we renovated each floor and started the concept for neighbourhoods." Research teams were brought together in an area containing labs, offices, meeting rooms, break areas and shared equipment. The plan was to make people feel at home in a large building.

Genentech Hall, which was designed by US architects SmithGroup, with the help of Zimmer Gunsul Frasca Partnership, based in Portland, Oregon, is home to more than 70 different labs. Each 'neighbourhood', housing four or five research groups, has experiment space on the outside and offices in the middle. Research teams within a neighbourhood may share common research interests — as well as equipment, meeting rooms and eating areas — but are not necessarily from the same departments.

The Ray and Maria Stata Center at the Massachusetts Institute of Technology (MIT) in Cambridge, set to open in winter next year, will also subscribe to the neighbourhood concept. The centre's researchers, spanning a diverse range of fields, from linguistics to artificial intelligence, will occupy a series of areas, each with a two-story lounge at its centre. The building's architect, Frank Gehry, devised the scheme to provide a visual and physical connection between two levels. In addition, each neighbourhood has a central elevator core, a central staircase and some double-height labs.

Proof of principle

But will such innovative designs really bear fruit? Is the extra cost of connecting walkways justified in terms of real scientific integration? And how do we know whether natural light and open spaces get the best out of researchers? For all the good intentions of architects and the scientists who advise them, there is no proof that any of these tactics actually helps to produce better science. "Generally speaking, architects have good intuitive beliefs and some anecdotal evidence, but little beyond that," says Kevin Kampschroer, director of research for the US General Services Administration (GSA), an organization based in Washington DC that provides and maintains workspace for federal employees.

The problem that is now beginning to be considered by architects. Norman Koonce, executive vice-president of the Washington-based American Institute of Architects (AIA), discussed the issue in an editorial in the spring 2003 issue of the AIA's *Journal of Architecture*, asking: "What would it mean for architects to move beyond an intuitive and anecdotal rationale in their design? How much better could we serve our clients and the public if we understood how their brains enable perception of their physical



New trends: glass walls (top) and open common areas (above) are recent themes in lab design.

environment and generate physiological responses to it?"

The San-Diego-based Academy of Neuroscience for Architecture may be the place where such questions can be answered. The academy was unveiled in May, and aims to support research that bridges the gap between neuroscience and architecture. One project that will soon be under way will use functional magnetic resonance imaging to identify the brain regions that are activated when people look at historic houses, schools and churches, as opposed to contemporary ones. Other research being considered by the academy includes how lighting and acoustics can affect cognitive activity, and how the brain enables people to find their way around in complex buildings. Although such issues are not specific to lab design, they should produce results about issues such as lighting that could potentially make labs better places to work.

Little is currently known about these issues, or even about how to study them. "We have no clue how design affects the nervous system," says Gage, who has been involved in establishing the academy's research direction, and will advise its researchers when they begin work. "We don't even know what are the right questions to ask." John Eberhard, AIA director for research planning and a key figure behind the academy, believes that it will be 10–20 years before neuroscientists know enough about how the brain functions to begin to address these questions. "We are just getting started," he says.

In the meantime, projects with less lofty

goals are under way. The GSA has entered a collaboration with the AIA and scientists at the US National Institutes of Health to look at the effects of stress on productivity in the office environment. Within the next month, the GSA will recruit about 200 study participants currently working in old-fashioned office environments, with features such as little natural light. The workers will spend time in three different office spaces over a period of two years, during which time their stress levels will be tracked by monitoring indicators such as heart rate and hormone concentrations.

The study will begin in the workers' existing offices, after which they will move to temporary accommodation, where they will continue to be monitored. Meanwhile, their offices will be renovated so that factors such as space, lighting, acoustics and temperature are altered. The final measurements will be taken back at the renovated offices. During this period, the workers will also be asked about how they feel about aspects of their new environments, such as lighting.

"What we are doing is unique," says Kampschroer. "There have been some studies that have looked at narrow aspects of one question or another, but nothing so comprehensive." He hopes that this study, and others planned by the GSA, will provide design guidelines based on empirical evidence. "We will be able to design for reducing stress," Kampschroer predicts.

Although researchers warn that they may not isolate a single variable that reduces stress for all workers, such projects could form the basis for a proper understanding of how workers interact with their environments. Further down the line, such studies could also contribute to a thorough evaluation of the way in which laboratories are designed, although this is still many years away. Eberhard draws comparison between architects today and medical doctors in the 1860s: "They did their best for patients but did it before anyone knew about mechanisms of disease. We think the same shift in knowledge that happened to medicine could happen to architecture."

Laura Bonetta is a freelance writer based in Bethesda, Maryland.

Academy of Neuroscience for Architecture

♦ www.neuroscienceforarchitecture.org

T-ray specs

Radiation from a previously unexploited region of the electromagnetic spectrum could hold the key to a new generation of security devices. Catherine Zandonella investigates.

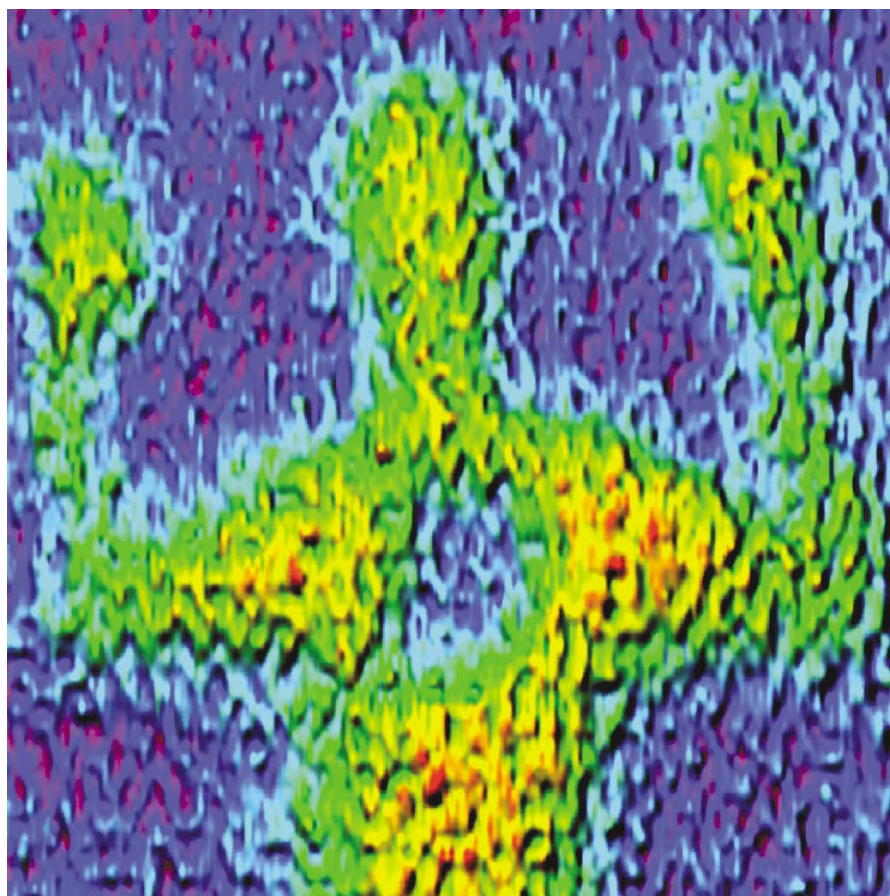
Suicide bombers, plastic explosives strapped to their bodies, approach the turnstiles at a packed football stadium. The security guards don't have time to search every spectator, and even if a metal detector were installed, it would miss the terrorists' deadly cargo. But a novel device that can see through the bombers' clothing succeeds where other systems fail. Security personnel are alerted, and surround the attackers before they can strike.

Such is the potential power of a new imaging technology. Terahertz devices, so named because they detect electromagnetic radiation in the terahertz frequency range (1 THz is 10^{12} Hz), promise to peer through clothing, revealing concealed weapons and explosives. The technique could also be used to seek out structural defects in materials, to detect skin cancer or to provide new information about astronomical objects.

For years, radiation with a frequency of between 0.1 THz and 10 THz languished unexplored. But recent advances in generating and analysing such radiation, together with an avalanche of research funding for antiterror applications, are now helping researchers to examine its applications.

Terahertz radiation, often called T-rays, lies between microwaves and infrared light in the electromagnetic spectrum. It can sail through paper and clothing, but not very far through skin or biological molecules, so objects hidden beneath clothing can be revealed by analysing the way in which the radiation is scattered and reflected off individuals. And because biological molecules absorb different frequencies of terahertz radiation, it may even be able to show whether a seemingly innocuous white powder is actually something more sinister such as anthrax.

That such goals seem achievable owes a lot to two landmark developments in terahertz research. One breakthrough came in the form of an imaging technique known as time-domain terahertz spectroscopy¹, which was developed at Bell Laboratories in Murray Hill, New Jersey, in the mid-1990s. Rapid bursts of visible light are fired at a semiconductor crystal, which absorbs the radiation and re-emits it as pulses of terahertz waves, which are then directed at the object to be imaged. Some of the radiation passes into the object and bounces off its internal layers. By scanning the pulses across the target and measuring the time it takes for the reflected



Caught: this terahertz image clearly shows an object (blue) concealed beneath the person's clothing.

radiation to reach a detector, researchers can build up a three-dimensional image of what is under the object's surface.

Light work

A more recent advance came from a team led by Alessandro Tredicucci, a photonics researcher at the National Enterprise for Nanoscience and Nanotechnology in Pisa, Italy. Terahertz research had been handicapped by the fact that the only lasers that could produce terahertz radiation were huge devices that filled buildings or which required extreme operating conditions. Tredicucci's team cracked the problem last May², when it produced a terahertz laser just a few millimetres in size.

Lasers use electricity to boost electrons to higher energy levels: the electrons emit light when they fall back to a lower level, and the frequency of that light depends on the gap between the levels. The energy gaps in various

materials are suitable for producing infrared or visible light, but Tredicucci's group, working with a team at the University of Cambridge led by Giles Davies, was the first to create a material that could generate terahertz radiation. They used a semiconductor crystal made from alternating layers of gallium arsenide and aluminium gallium arsenide. The layers allow electrons to cascade through a series of energy levels, emitting a photon each time to give a 'quantum cascade' terahertz laser.

Xi-Cheng Zhang, a pioneer of terahertz imaging at Rensselaer Polytechnic Institute in New York, is one of the researchers capitalizing on such advances. He was approached in April by Lockheed Martin Space Systems in New Orleans, which made the insulating foam implicated in the space shuttle Columbia disaster. One of the shuttle's wings was damaged during take-off by a chunk of foam that fell off the fuel tank, and this is believed to have caused the craft's destruction as it

returned to Earth. Lockheed wanted to see whether terahertz imaging could check for defects in the foam before the launch of future missions.

The company sent Zhang a block of foam with four holes hidden in it, the smallest just 6 millimetres across. Using time-domain spectroscopy, Zhang's team monitored the reflection of terahertz waves from inside the foam and spotted all four holes. The group is now running a feasibility study to see whether the technology can be used to test all of the foam used on a shuttle's tanks.

Zhang's team is also working on security applications. Shortly after the 2001 US anthrax attacks, during which envelopes containing anthrax were posted to politicians and media outlets, Zhang and his colleagues began investigating whether terahertz radiation could help to prevent similar events in the future. They used envelopes containing various white powders — flour, sugar, talcum powder and spores of a benign species of bacterium, which acted as a surrogate for anthrax — and found that they could detect a characteristic absorption signature for the spores³. Several groups, including a team led by Daniel van der Weide, an electrical engineer at the University of Wisconsin in Madison, are doing similar work in an attempt to develop a mail sorter that can check for biological hazards.

Snap happy

But firing terahertz radiation at an object is not the only way to get an image. All objects naturally emit the radiation and so can be imaged using an appropriate detector. Last autumn, an international consortium of industrial and academic researchers known as StarTiger built the first such device. The work was sponsored by the European Space Agency, which hopes to use StarTiger's camera to study terahertz radiation emitted by astronomical objects such as gas clouds.

The camera's potential security applications are now being explored by researchers at the Rutherford Appleton Laboratory near Oxford, UK, where StarTiger was hosted. Scientists there believe that the camera will detect concealed objects such as plastic explosives from tens of metres away. In a feasibility test, they took images of people last month (see previous page), and they are now working on increasing the camera's resolution to an appropriate level for practical applications.

Such work has caught the attention of the US military. The Army Research Office in Research Triangle Park, North Carolina, is funding research aimed at producing similar devices. But, unlike the Rutherford Appleton team, the military argues that relying on naturally emitted radiation does not give a strong enough signal, so it wants its devices to



What lies within: researchers have successfully used terahertz radiation to spot internal defects in insulating foam (inset) for the space shuttle.

others are trying to make a laser whose frequency can be tuned across a limited frequency range. Another problem that the researchers face is that the lasers need to be cooled to below 100 kelvin using liquid nitrogen — a cooling system that cannot fit on a hand-held device. This is not a drawback for all applications — a mail-sorting device, for example, would not need to be portable.

But the detection of biological agents also carries a risk that scattering could be mistaken for absorption. Particles that are roughly the same size as the wavelength of terahertz radiation, such as fine sand about 1 mm across, will scatter specific frequencies, creating a signal similar to an absorption spectrum. In addition, there is the question of safety. Researchers believe that their devices are safe, as they produce radiation at power levels well below those used by mobile phones, but only limited work has been done on the effect of terahertz radiation on living tissue.

Although it seems that terahertz technology has a number of unresolved issues, researchers are not particularly concerned. "The field is in its adolescence," says Roger Appleby, a physicist with defence company QinetiQ in Malvern, UK. "It is unfair to compare terahertz to more mature fields."

If Appleby's confidence is justified, terahertz radiation could go from being a backwater of the electromagnetic spectrum to a region that makes possible a new generation of security devices. Tredicucci, for one, is in no doubt. "If you need to detect something hidden," he says, "this is the region with which to do it."

Catherine Zandonella is a freelance writer in New York.

include their own terahertz radiation emitter.

Generating sufficient power to illuminate a target is difficult, and current devices are a long way off this goal. Terahertz waves below 0.5 THz are used, as higher frequencies are absorbed by water vapour and can't travel more than about a metre through air. To produce these signals, physicists use a microwave emitter and convert this radiation into T-rays with frequency multipliers, which insert extra cycles into the chain of electromagnetic waves. But these multipliers are very inefficient, so it is difficult to get sufficient power for a clear image. Van der Weide believes that by improving the multipliers and the antennas that emit the radiation, he will be able to make a device that can image hidden objects, such as landmines, from 10 metres away with greater resolution than the Rutherford group's camera can achieve.

How low can you go?

For detecting biological molecules, the relative absorption of terahertz radiation is key. But to test for a wide range of materials, you need a laser that emits across a broad range of frequencies. Although Tredicucci's laser is promising, it can only operate at a fixed frequency. Initially this was 4.4 THz, but his team, and other groups, have recently made lasers that go as low as 2.5 THz, and they are confident of pushing the frequency lower.

These lasers could be used to seek out biomolecules that absorb terahertz radiation at specific frequencies, although Tredicucci and

1. Hu, B. B. & Nuss, M. C. *Opt. Lett.* **20**, 1716–1718 (1995).
2. Köhler, R. *et al. Nature* **417**, 156–159 (2002).
3. Wang, S., Ferguson, B., Abbott, D. & Zhang, X.-C. *J. Biol. Phys.* **29**, 247–256 (2003).

Lack of concern deepens the oceans' problems

Research has exposed flaws in marine conservation, yet little is written and less is cited.

Sir — There is a crisis in the world's oceans^{1–4}. Myriad human impacts — from coastal development, pollution and habitat alteration to introduction of invasive species, overfishing and climate change — seriously jeopardize marine ecosystems and the services they provide. How much research is being done in these crucial areas?

Earlier analyses in the period up to 1996 documented that marine ecosystems were represented by only around 5% of papers in two main applied-ecology journals^{5,6}. We have discovered that there is still, eight years later, negligible emphasis on marine issues in the conservation-biology literature and that conservation receives very limited attention in the general marine and fisheries literature. For example, only about 5% of papers in leading marine-ecological journals deal with pollution, nonindigenous species, overfishing or marine protected areas.

We next asked if the impact of terrestrial

papers published in *Conservation Biology* differed from the impact of marine papers. To do this, we looked up every research paper published in *Conservation Biology* in 1997 in ISI's Web of Science and tallied the number of citations each paper received (we chose 1997 to allow time for papers to be assimilated into the literature). Papers on marine topics were cited an average of 7.1 (s.d. 3.5) times, whereas terrestrial articles were cited an average 18.2 times (s.d. 14.6). Thus, not only were fewer articles being published about marine conservation, but those that were published appeared to have less impact than research on terrestrial habitats.

Most attention to conservation issues has focused on the terrestrial realm, but it is clear that the oceans face an increasingly severe array of problems. Humans might well focus first on the habitat they themselves inhabit, but it is clearly time for

more attention and resources to be directed towards the oceans. The limited rigorous scientific research that has been done has exposed many of the flaws in our present management paradigms⁷. Our ability to overcome the problems will surely depend on the contributions of science.

Beth F. Kochin, Phillip S. Levin

*Northwest Fisheries Science Center,
National Marine Fisheries Service,
Seattle, Washington 98112, USA*

*Literature and citation analyses available direct
from P.S.L. at phil.levin@noaa.gov*

1. Myers, R. A. & Worm, B. *Nature* **423**, 280–283 (2003).
2. Baum, J. K. et al. *Science* **299**, 389–392 (2003).
3. Jackson, J. B. C. et al. *Science* **293**, 629–638 (2001).
4. Pew Oceans Commission. *America's Living Oceans: Charting a course for sea change* (Pew Oceans Commission, Arlington, Virginia, 2003).
5. Irish, K. E. & Norse, E. A. *Conserv. Biol.* **10**, 680 (1996).
6. Ormerod, S. J. *J. Appl. Ecol.* **40**, 204–213 (2003).
7. Peterson, C. H. & Estes, J. A. in *Marine Community Ecology* (eds Bertness, M. D. et al.) 469–509 (Sinauer Associates, Sunderland, Massachusetts, 2001).

Impact factors: just part of a research treadmill

Sir — In Brazil, scientists need to publish more every year to obtain scarce funds, leading to an exaggerated degree of competitiveness and promoting a cultural distortion where scientometrics prevails over knowledge (as discussed by P. A. Lawrence, *Nature* **422**, 259–261; 2003 and in subsequent Correspondence). In this highly competitive atmosphere, Brazilian science is improving. But what is the impact on the individual, particularly on new PhDs who need to establish a career?

In 2001, the Ministry of Science and Technology created a programme to affiliate 10 promising young PhDs with research centres. There were 1,154 candidates for these 10 positions. In the same year, the National Research Council offered two-year research grants varying from US\$2,000 to \$43,600. For the section that included biochemistry, biophysics, physiology, pharmacology and neuroscience there were 437 applications, of which 267 were approved on merit but only 20 were funded. The main selection criterion for both programmes was based primarily on number of publications and impact of the journals concerned.

We interviewed postgraduate students, postdoctoral fellows and teaching staff in one department about their concerns. A common feature was a high degree of involvement with their work. A typical statement by a postgraduate student

referring to the faculty was: “They must be crazy. They live, they eat, they will probably die in that laboratory. They arrive at 8am and never leave before 10pm. Why do you think they marry among themselves?”. (The 39 established investigators included 13 married couples.)

On publication pressures, typical statements were: “[The adviser] doesn't care about my thesis as such. He believes that a thesis is the consequence of good work and good work means papers published in good journals.” Referring to colleagues' work, people mention the number of publications and the journal, not knowing exactly what had been discovered. “They evaluate people by the number of publications ... and classify them by the impact of the journals: high-impact and low-impact scientists.” Or: “If you publish a paper in *Nature*, marvellous, but if you do it in a Brazilian journal they will say, ‘Look what a lousy contribution to science.’” Submitting a paper evokes strong emotions: “When the journal does not accept ... you feel as if it is not only your paper but you yourself that is rejected”. Or: “It is a great feeling to have a paper accepted ... when you know that so many people have their papers rejected.”

The difficulty of obtaining research support, in a country where funding is mainly public, generates strong feelings of insecurity at all levels. “You never know if you will have money, if your application is going to be approved.” Or: “If you stop publishing you lose your grant ... You are ejected from the system; it doesn't matter

what you did in the past — it only matters what you have done in the last two to three years.” One respondent said: “I knew that they would post the result of the evaluation. ... I went to the computer. My heart was pounding inside me ... My name was there, I was awarded ... I started crying and I couldn't stop. ... I went and hugged my wife, crying ... and all of that for a lousy grant of less than US\$8,000 per year.”

In universities where there is no research, the thesis defence is a rite of passage, legitimization as part of the teaching staff coming mainly from the academic title of PhD. For a research postdoc, the analogy is the publication of a paper, but this affords only temporary respite, not a transition to a ‘new world’. The doctoral thesis in itself is unimportant; what counts are ‘papers published in good journals’. According to our interviews, legitimization never really arrives. The trajectory of the scientist becomes an increasingly difficult struggle for grants, where the individual may lose support at any time. The idea of a continuous, stable career is blurred. Instead people are in perpetual transition, repeatedly having to prove their capability, and at increased risk each year of either being eliminated or burning out if they remain in the system.

Leopoldo de Meis*, Maria Scarlet do Carmo*, Carla de Meis†

**Departamento de Bioquímica Médica, Centro de Ciências da Saúde, Universidade Federal do Rio de Janeiro, Rio de Janeiro 21941-590, Brasil*

†Instituto de Psiquiatria, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 21941 590, Brasil

A stranger in a strange land

Bill Bryson's latest journey takes him through the unfamiliar world of science.

A Short History of Nearly Everything

by Bill Bryson

Doubleday: 2003. 528 pp. £20

Broadway: 2003. \$27.50

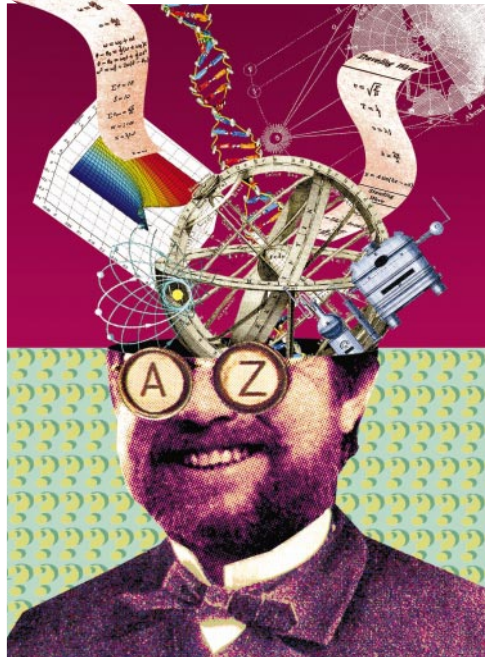
Walter Gratzner

This is what Bertrand Russell wrote about his omnivorous friend, Aldous Huxley: "You could tell by his conversation which volume of the *Encyclopaedia Britannica* he'd been reading. One day it would be Alps, Andes and Apennines, and the next it would be the Himalayas and the Hippocratic Oath." A similar, if a little more circumscribed, social accomplishment can be yours if you last the pace with the irrepressible Bill Bryson in this canter through the landscape of science and its historical roots.

Let it be said at once that Bryson's is a truly remarkable achievement. He was awakened, as he tells it, from a state of unblemished and tranquil ignorance, while looking out of an aeroplane window, by a spasm of curiosity about the nature of the world beneath, not to mention the skies above and the Universe beyond. From this damascene moment sprang the extravagant ambition to acquaint himself with the entirety of science and interpret it to the citizenry at large. He accordingly immersed himself, if his bibliography is anything to go by, in a formidable course of study, but more than that, he approached many scientists, hooked his finger in their buttonholes and did not let go until he had got from them all he wanted. Bryson was surprised and delighted to find that most of them were willing, even eager, to talk to him at length about their work. His quest took him three years.

As aficionados of his travel writing will know, Bryson is quick — some might say profligate — with the catchy trope and the arresting analogy. So, the diameter of an atom is to a millimetre as a sheet of paper is to the height of the Empire State Building; if the Earth were an apple, the deepest holes yet sunk by geologists would still not begin to penetrate the skin; and if you could retreat through time at the rate of one year per second it would take you three weeks to arrive at the emergence of humans, but twenty years to reach the Cambrian, the period of the Burgess Shale fossils. (Do I hear the creak of mouths falling open?)

It is easy to discern where Bryson's enthusiasms have chiefly settled. He is best on biology, especially evolution and palaeontology, and also on geology and on the cosmos. He waxes lyrical on the numberless riches in the caverns of the Natural History Museum



in London, perhaps because he found there a genial host, the matchless Richard Fortey, author of captivating books on trilobites and the history of life on Earth. Fortey guides him through labyrinthine disputes about events that happened 500 million years ago, and through the ill-natured conflicts between the rival evolutionists Stephen Jay Gould, Richard Dawkins and Simon Conway Morris, and the gaffes that characterized the interpretation of the Burgess Shale fossils.

Bryson laments with Fortey the fading of classical taxonomy. There are said to be fewer than 20 taxonomists in Britain now to patrol an endless frontier, with some 15,000 previously unrecognized species surfacing each year in the literature; in the nineteenth century — the golden age of taxonomy — they would all have been subjected to minute scrutiny, sometimes close on a lifetime's work. In all, the proportion of the planet's species so far recognized is (depending on whom you believe) somewhere between 2% and 12% (and in the case of bacteria and viruses, well below 1%). It is sad when the world's only expert on a genus of bryophytes, say, heads for retirement or the grave — he is irreplaceable. Yet Bryson does not altogether explain why the effulgent new layer of understanding exposed to view by genomic analysis might not be considered more exciting and more pressing than classical morphology.

Bryson's grasp is somewhat less secure when it comes to the basics of the physical sciences. He relates the circumstances of

Lavoisier's career and death on the scaffold, and reveals that the lineaments of a statue erected in his honour a century later were in fact those of another victim of the Terror, Condorcet, recycled by the sculptor. This is diverting enough, but tells us little about the unparalleled magnitude of Lavoisier's achievements. The egregious Count Rumford gets a walk-on part, and it is pleasing to learn that he achieved immortality by inventing thermal underwear and the drip coffee-maker, but what about his observations in the cannon-boring works, which disposed of the caloric theory of heat, just as Lavoisier consigned phlogiston to oblivion?

There are, inevitably in a work of such scope, some errors of fact, and (more seriously) a few of comprehension. What is one to make of the assertion that "oxygen itself is not combustible", whereas hydrogen gas "is extremely combustible"? Or that in outer space, "each molecule is very

warm". And when the sun is hot on your back, Bryson says it's the impact of "excited atoms" that you are sensing — so what of radiant heat?

But one can readily forgive a man such lapses who can sum up the state of cosmology thus: "The upshot of all this is that we live in a universe whose age we can't quite compute, surrounded by stars whose distances we don't altogether know, filled with matter we can't identify, operating in conformance with physical laws whose properties we don't truly understand."

Bryson's brief valedictory chapter is about what humans are doing to our planet. Over the aeons of biological time, he tells us, one life form, on average, vanished every four years; now the rate of extinction of species resulting from human activities exceeds this by a factor of 120,000. Man, as Julian Huxley expressed it, has become the cancer of the Earth.

The last man who was said to know the whole of science was the Victorian savant William Whewell. Bryson would make no such claim for himself, and indeed is modest about his learning, but he has shown that science — most of it, at any rate — is accessible to an outsider equipped with a decent measure of curiosity and some staying power. His is a bravura performance, which will give much pleasure, and I salute it.

Walter Gratzner is emeritus professor of biophysical chemistry, King's College London, the Randall Centre, Guy's Campus, St Thomas' Street, London SE1 1UL, UK.

Living with the Neanderthals

Darwin's Children: Evolution Has Changed the Face of the World

by Greg Bear

HarperCollins: 2003. 400 pp. £17.99

Del Rey: 2003. 368 pp. \$24.95

Michael A. Goldman

Science, the ultimate arbiter of truth, is still stained by the imperfections of human nature. We sometimes think we have all the answers. But the preposterous ideas of yesterday are the unshakeable dogmas of today, and the ancient superstitions of tomorrow. Science is driven by politics, and politics by fear. Greg Bear's latest novel, *Darwin's Children*, grapples with the biases of our society, and the machinations of science, politics and science policy, in the face of its most formidable challenge for more than 25,000 years: the next step in human evolution. Today's wild speculations might not make it into tomorrow's textbooks, but Bear's ability to tell a good story is surpassed only by his enthusiasm for the advancing edge of molecular biology.

In *Darwin's Radio*, the forerunner to *Darwin's Children*, fictional molecular biologist Kaye Lang and anthropologist Mitch Rafelson were at the epicentre of a radical new view of 'human' genetics and evolution. Human endogenous retroviruses (HERVs) had suddenly mobilized and were wreaking havoc on the newly sequenced genome. Herod's plague, later dubbed SHEVA (for scattered human endogenous retrovirus activation), infected pregnant women, causing spontaneous abortion of a fetus and subsequent development of a 'daughter' zygote that had undergone virus-directed mutation. The children born of SHEVA pregnancies represented the next step in human evolution. They communicated using a combination of double-speak, olfaction and skin freckling.

The plague had been seen at least twice before — once at the scene of a massacre of pregnant women in the recent past, and once in the Ice Age, when a late-stage Neanderthal couple gave birth to an early-stage modern human. The mobilization of the endogenous retroviruses, it seemed, was a response to stress that shuffled the genome in a relatively organized way, and potentiated speciation, or at least subspeciation. Predictably, Rafelson and Lang had trouble convincing the scientific establishment that we weren't just dealing with a standard epidemic, but a natural and desirable advance.

At the opening of *Darwin's Children*, modern society isn't ready to share its bounty with a new race of human, or a new species of hominid. The 2.5 million surviving 'virus

children', or Shivites, are mostly confined to special schools or concentration camps. New branches of law enforcement and public health, enjoying the federal budget dividends of fear, have been established to deal with the problem. Rafelson and Lang themselves had a SHEVA child, Stella Nova, whom they raised out of the grip of the government's Emergency Action directives, and did their loving best to conquer a quantum generation gap. SHEVA dominates the political scene, shifting our way of life. Habeas corpus is suspended for renegade Shivites, as it was in the aftermath of the September 2001 terrorist attacks for Middle Eastern men suspected of terrorist links. The world is gripped by a dread of nature, much as we are today by diseases such as SARS or AIDS. Viruses jump species barriers as humans tamper with the balance of nature.

In weaving his story, Bear takes us through dozens of enticing didactic excursions filled with speculation about everything from molecular biology to God. Writes Kaye Lang: "The role of SHEVA in the production of a new subspecies is but one function... They are mediators and messengers between cells, ferrying genes and coded data around many parts of the body, and even between individuals... They are like original sin." Peripheral ideas abound, but not so very long ago the concept of a simple molecule such as DNA being an information-carrying instrument was inconceivable.

The epiphany scenes, in which super-heroine Lang meets an all-loving, all-forgiving "caller", will shock the most die-hard science-fiction fans, and stun the average *Nature* reader. Bear does an admirable job of trying to square religious beliefs and scientific

reason when functional magnetic resonance imaging is used to see what a revelation looks like. And he never takes us quite to the brink of a miracle, stating emphatically in an epilogue that he isn't advocating special creation or God-directed evolution. In short, he reminds us that the mystic experiences that people have are neurological events as real as any other perception, and suggests that they might just be manifestations of a higher organic (as opposed to spiritual) existence.

Like all the best science fiction, *Darwin's Children* is a medium for telling the public about science. Through one of his characters, Bear pleads for biologists to write clearly: "Don't they realize how important it is to get the word out to everybody?"

Bear makes a few mistakes of definition, going beyond the studied speculation expected in a hard science-fiction novel. For instance, he makes reference to the odd karyotype of a virus child as "52 xx" — convention would have it written as "(52,XX)". He calls this a polyploid number, but it isn't a multiple of the haploid human chromosome number of 23, so really it should be 'aneuploid'. Bear incorporates the new 'fact' that humans have some 30,000 genes into his definition of 'genome', accepting it as canon, but he rejects the less controversial (albeit possibly incorrect) 'fact' that we are not descended from the Neanderthals.

Although *Darwin's Children* works well on its own, readers who skip *Darwin's Radio* and plunge straight in will miss some of the intense excitement of mystery and discovery. *Darwin's Children* is more a chronicle of the new humanity's first attempt to share the planet with its forebears, and contains a lot of good expository prose about the

Art

Seeing the light

Light may be the life force of fine art, but light itself takes centre stage in Hiro Yamagata's imposing exhibition in the Japanese seaport of Yokohama.

In a display co-hosted by NASA, a pair of 20-metre cubes covered with holographic panels (right) make art out of sunlight, refracting its beams to shower coloured rays over surrounding buildings.

Nearby, Yamagata's dizzying NGC6093 installation, which is named after a dense star cluster in the constellation Scorpio, consists of an enclosed room with holographic walls and floor, filled with dangling mirror cubes. Rotating lasers bounce beams through this maze of light sources. In one of the most beautiful — and unintentional — effects, green light catches the nylon that is used to hang the mirror cubes, and pours down like electric rain.



Yamagata says that his laser shows mark a deliberate attempt to distance himself from the commercial paintings for which he is best known in Japan.

David Cyranoski

"The World of Hiro Yamagata and NASA" can be seen at Yokohama, Japan, until the end of September.

www.hiroyamagata.net

virus children's relationship to their world.

Bear's two Darwin novels were not written just to entertain. He also seeks to teach readers about science, to highlight our utilitarian politics and our inability to get along with each other, and to provide a quasi-rational basis for theology and morality. He advances a world view in which we are all part of the vast neural network of life, cutting across ethnic borders, species divides and the chasms between taxonomic kingdoms, in balance, and in two-way communication, with the ecosystem. Bear might be off the mark, or he might just be anticipating the next giant leap in our understanding of evolution and ourselves. ■

Michael A. Goldman is in the Department of Biology, San Francisco State University, San Francisco, California 94132-1722, USA.

The last of its kind?

*Pheidole in the New World:
A Dominant, Hyperdiverse
Ant Genus*

by Edward O. Wilson

Harvard University Press: 2003. 818 pp.
\$125.00, £83.50, €125.00

Donat Agosti

"Everything goes." Philosopher Paul Feyerabend's statement of anarchy in scientific method — that you can research however you like — might well hold for taxonomic monographing. The field is plagued by an increasing shortage of specialists, despite the ever-growing pressure to complete the 'catalogue of life' with descriptions of new species and revisions of known species. *Pheidole in the New World* makes a sizeable advance, describing 624 species of ant, of which 337 are new to science. It was written by Harvard University biologist E. O. Wilson, one of the world's leading biologists and environmentalists, and published by Harvard University Press, which has a much higher profile than most of the 500-plus journals in which ant descriptions usually appear.

This important book is modelled on the highly successful Peterson's field guides. Each species is given one page, of which half is covered by lavish line drawings with very helpful indicators marking the diagnostic characters. The rest of the page contains a short description (including a series of standard measurements), a handy indication of closely related species, a summary of key biological features and the distribution range. Unfortunately, details of the biological materials examined are either missing or incomplete.

The book's introduction includes some notes on the ecology of the genus *Pheidole*, some generalities on hyperdiversity, an illus-



Adding ants: *Pheidole subsphaerica* (top), *P. cursor* and *P. lupus* are three of 337 new ant species described by E. O. Wilson.

trated glossary, and an uninspiring, unillustrated key to species. The *pièce de résistance* is an accompanying CD-ROM, which comprises fantastic colour images of about half of the species and a simple search engine.

Ant taxonomy is at a vibrant stage of development, with more than 130 active scientific contributors, all but two of whom are connected over the Internet. Since the last catalogue of the world's ants was published — Barry Bolton's *A New General Catalogue of the Ants of the World* (Harvard University Press, 1995) — 1,630 new ant species, including those in this book, have been described. A database (www.antbase.org) with a single web page for each species provides links to an increasing number of extraordinary websites, organized by region (such as www.antweb.org) or theme (GenBank).

The entire 3,500 taxonomic publications covering the non-copyrighted systematics of the 11,574 known species of ant will be

available online this autumn, providing an enormous free resource. Wilson's *Pheidole in the New World* will not be part of this system, as the book is copyrighted: as a result, apart from the unprotected nomenclatorial information and links to those images available at Harvard's online Primary Type Specimen Database, its 624 descriptions cannot be readily integrated into the world-wide biodiversity knowledge base. This raises serious issues about an academic publisher's obligations to make biodiversity data freely available for non-commercial research, education and conservation.

The inaccessibility of published records frustrates efforts to compile lists of the world's living species, and to develop the next generation of tools and data repositories that can be seamlessly woven into systems such as GenBank, biodiversity monitoring systems, or the fledgling Global Biodiversity Information Facility. Records are hidden in thousands of journals and books, and access to new work is increasingly restricted by an ever more proprietary copyright environment. Money aside, even password protection impedes the building of information networks.

Most of the species described in *Pheidole in the New World* are based on material collected in the tropics, where most of the world's species live and vanish. The Convention on Biological Diversity, sadly still not ratified by the United States, calls for the exchange of scientific information and repatriation of data to the place of origin. Selling expensive books doesn't aid this process.

Wilson isn't just a taxonomist — he is a prominent biologist and an outspoken environmentalist. His style of science and his involvement with conservation set precedents far beyond taxonomy. He espouses the development of an Internet-based 'encyclopaedia of life', stresses the virtue of the biodiversity commons — a movement to make biodiversity information freely accessible for non-commercial use — and would like to see taxonomy follow the DNA revolution (*Trends Ecol. Evol.* **18**, 77–80; 2003), but his book on *Pheidole* doesn't reflect these ideals.

In these days of biodiversity crises, taxonomy isn't just a science where "everything goes". It has the potential to play a pivotal role in quantifying changes in biodiversity. It is not enough to talk about the biodiversity crisis; action is required. It is a pity that Wilson, with his ingenuity and access to resources, did not grasp the opportunity to present these important data in a more novel and useful way, but instead preferred to produce, in his own words, the "last of its kind": another visually appealing monograph. ■

Donat Agosti is in the Division of Invertebrate Zoology, American Museum of Natural History, Central Park West at 79th Street, New York, New York 10024-5192, USA.

Conscious efforts

Summarize yourself in the form of a title of a paper in Nature.

A fisher in the sea of neuroscience.

What was your first experiment as a child?

It was an attempt to describe the speed of fishes in an aquarium by drawing a line on paper alongside, at the same speed.

Who has been the most important mentor in your career?

Friedrich Nietzsche, when he likened scientists to prehistoric hunters. Our modern scientific activities are driven by hunting instincts that we have inherited from our ancestors.

What single scientific paper or talk changed your career path?

The Nobel prizewinning papers by Alan Hodgkin and Andrew Huxley (*J. Physiol. (Lond.)* **117**, 500–544; 1952), in which they described nerve impulses in terms of ionic conductances, remain unsurpassed examples of beauty in science.

What's your favourite conference destination, and why?

I prefer to go to a new place each time, to avoid the risk of spoiling pleasant memories.

What book is currently on your bedside table?

The Illusion of Conscious Will by Daniel M. Wegner.

Assuming the dead can be raised and/or time travel exists, who from the world outside science would you most like to have dinner with?

Lev Tolstoy — and on such an occasion I would stay vegetarian, of course.

You are on a plane behind two students obviously going to the same conference, who start to talk about your work. What do you do?

I was in a still more frustrating situation at a meeting, when some smart-alec asked my student whether Oleg Krishtal was still alive. I could not stand the pleasure of immediate resurrection.

What one thing would you rescue from your burning laboratory?

My copy of *Ion Channels of Excitable Membranes* by Bertil Hille, with a personal inscription by the author.

What's the best piece of advice you've received?

Keep cool!

What do you most dislike about having research published?

I quote the Russian poet Tyutchev: "A thought, once uttered, is a lie."

The Internet is the bane of scientists' lives because...

...it's a sexton of personality. Still, we shall have a lot of entertainment on the way.

What do you do to relax?

Reflect.

What's just around the corner?

An intellectual continuum whose disciplinary borders have been crushed, like the Berlin Wall.

Is there a 'tyranny of reductionism' in how scientists are trained today?

I consider reductionism as a good basis for any scientific training. The innate human drive to metaphysics has to be properly balanced.

What overlooked or underrated discovery really changed the science in which you work?

There is a story about the composer Mendelssohn, who bought a piece of herring from a market. When he got home, he found that it had been wrapped in a piece of sheet-music manuscript. Mendelssohn started to read and was overwhelmed — the music turned out to be a forgotten piece by an obscure composer called J.S. Bach. This discovery led to Bach's contemporary rehabilitation. With so many people involved in science, we are unlikely to find such forgotten gems. If something important has been overlooked, it will be reinvented at a price of losing the original reference.

Which field in science (apart from your own) deserves more funding, and why?

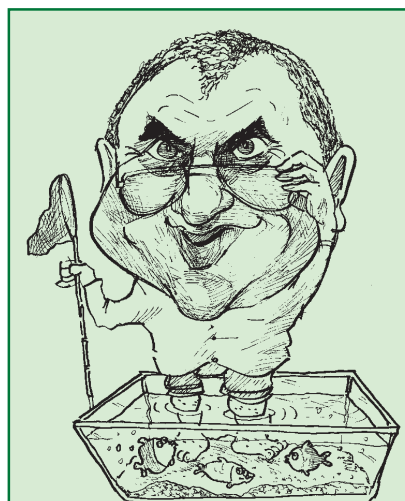
Not a field, but an aim: the efficient conversion of solar energy, either through solid-state technologies or the genetic engineering of novel plants.

What's the one thing about science that you wish the public understood better?

I recall a story in which people whose brains were being examined by positron emission tomography were asked to look at a blank screen. An image was gradually introduced by increasing the contrast, and the subject was asked to record the instant when an image was visible. It seems that the visual cortex detects significantly weaker signals than the conscious self. The public should better understand that science is not just about knowledge, but also self-knowledge.

What would you have become, if not a scientist?

I would be exploring the only thing that is not still subject to reductionist science — consciousness. The only available tool for the investigation of consciousness is reflection.



Oleg Krishtal

Oleg Krishtal is professor and chair of the department of cellular membranology at the Bogomoletz Institute of Physiology, Ukrainian Academy of Sciences. His research is in the fields of neuroscience, biophysics and pharmacology.

I have written a couple of books on this topic, published in Russian. One of them, a novel called *Homunculus*, is about a neuroscientist who discovers that his girlfriend is a smarter scientist than he is. This has been published in Russian and also in French under the title *Moi et Mon Double*. Another book, *To the Singing of Birds*, is a revelation of conscious self experiencing the times of prevailing darwinism. At the risk of seeming immodest, it awaits another Mendelssohn.

Do you have any more novels up your sleeve?

So far, only commentaries to the last one: it took me nearly two years to assess what it is that has come to my mind.

What one thing would you change about Nature?

I would add a section called 'Titles of rejected Letters' (peer-reviewed, of course).

What is the most interesting thing in your fridge?

Post-Communist non-filtered Ukrainian beer.

You've just been told (in confidence) that the world will end tomorrow. What would you do?

Relax.

Do you have a burning ambition to do or learn something of no practical or immediate value? If so, what?

I am afraid that life is a story of getting used to having only practical ambitions.

How would you like to be remembered?

As an optimist.

New player in pain

Edwin W. McCleskey

Long-term pain must be accompanied by long-term alterations in neuronal signalling — but what causes the changes? In rats, immune-like spinal cells are implicated, together with a molecule new to the field of pain.

Normally essential to warn us of danger, pain becomes harmful when it occurs without any evident stimulus. If this persists, it can cripple a person's life. A common kind of chronic pain, called neuropathic pain, often develops when nerves are damaged through surgery, compression by bone, diabetes or infection¹. Most research on the subject parallels the study of memory, seeking to understand how the communication between nerve cells — in this case, the neurons involved in transmitting pain signals — could show long-lived enhancement². A minority view focuses on cells of the immune system and related glial cells of the nervous system, proposing that they influence the connections between nerves³. That view now gains considerable support. On page 778 of this issue, Tsuda and colleagues⁴ show that spinal microglial cells use the P2X₄ receptor — a molecule with no previously known connection to the field — to induce a form of neuropathic pain. The importance of glial biology in chronic pain now seems undeniable.

Pain is sensed by 'nociceptive' neurons in the periphery of the body, which carry signals into the spinal cord; there, the nociceptive cells form connections (synapses) with spinal neurons (Fig. 1, inset). Some of these spinal neurons extend into the brain, while others form local circuits within the spinal cord. The first set of synapses in the pain pathway is considered the most significant site for 'gating' pain⁵; for example, opiates diminish the activity of such synapses, explaining much of the ability of these drugs to reduce pain. Conversely, if the synapses become hypersensitive, pain should increase. So the major goal of basic research into chronic pain is to understand how the connection between nociceptive and spinal neurons can become overly sensitized.

In their studies, Tsuda *et al.*⁴ used a classic model of neuropathic pain⁶: they cut a peripheral sensory nerve in a rat. A week later, a light touch caused the rat to withdraw its paw, as if from a highly unpleasant stimulus. Such hypersensitivity to touch is called mechanical allodynia, and is a serious form of neuropathic pain; in humans, it can mean that simply putting on a shoe is excruciating.

Tsuda *et al.* next looked at whether the allodynia was reduced by infusing P2X-receptor blockers into the spinal cord. P2X

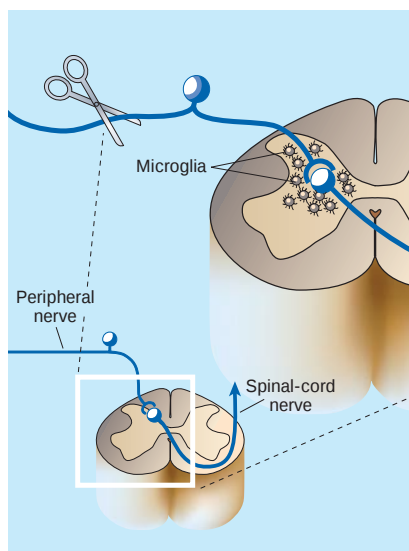


Figure 1 Pain and its pathways. Inset, 'nociceptive' nerve cells in the periphery of the body sense pain and send signals to spinal neurons that project to higher brain centres. Main picture, Tsuda *et al.*⁴ have found that cutting peripheral sensory nerves in rats causes microglial cells to congregate, and to produce P2X₄ receptors, on the side of the spinal cord where the cut neurons enter. The resulting chronic pain is suppressed by inhibiting P2X₄ receptors, and is mimicked by infusing microglia that bear activated P2X₄ receptors into the spinal cord of normal rats.

receptors are ion channels that sit in the outer membrane of certain cells; they open when ATP molecules bind to them from the outside of the cell, and allow Na⁺ and Ca²⁺ ions in⁷. The resulting alteration in ionic balance helps to control cellular activity. Of the seven subtypes of P2X receptor, two (P2X₃ and P2X₂) occur in large numbers on specific nociceptive neurons. So it is logical to test whether any particular kind of pain responds to drugs that affect P2X receptors. But Tsuda and colleagues' results seemed illogical. Allodynia was not prevented by a blocker called PPADS, which inhibits the P2X receptors on nociceptive neurons. But it was blocked by TNP-ATP, an inactive ATP mimic, at a high concentration that inhibits all P2X receptors.

So which P2X receptor is being blocked — and where is it? Tsuda *et al.* found that a fluorescently tagged antibody that recognizes

the P2X₄ receptor, one of the two subtypes that are insensitive to PPADS, produced strong fluorescence in the spinal cord of allodynic animals, and only on the side where the cut peripheral nerve entered. The antibody rigorously co-localized with a marker of microglial cells, and was absent from neurons. Microglia are a subset of glial cells; they are found in the central nervous system, and are known to synthesize just two P2X receptors, P2X₄ and P2X₇. Of these, only the P2X₄ receptor is insensitive to PPADS. As a final test of receptor identity, the authors perfused the animals' spinal cord with antisense DNA for the P2X₄ receptor — which prevents the receptor gene from being expressed — in the days following nerve damage. Mechanical allodynia was diminished.

Tsuda and colleagues' work is significant for its discovery that the P2X₄ receptor is involved in pain signalling. But perhaps its broader importance is the clear demonstration of a function for microglial cells. The authors elegantly tested this idea by injecting microglia directly into the spinal cord of undamaged rats. This by itself induced allodynia — but only if the microglia had first been incubated with ATP, thereby (presumably) activating them. A role for glia supports work by Watkins and colleagues; most recently, they showed⁸ that infusing general glial inhibitors near the spinal cord fully reversed the allodynia induced by inflammation of a peripheral nerve.

The term 'glia', which means glue, would seem to relegate these cells to being mere passive support for neurons. But microglia are hardly bystanders; rather, they behave like immune-system cells⁹. They migrate up gradients of attractive chemicals, congregating at sites of damage. And they can change from resting to active states. When active, they produce surface proteins that are characteristic of immune cells, and engulf cellular debris. Active microglia also secrete signalling molecules, including interleukins, tumour-necrosis factor and nitric oxide⁹, suggesting that they could alter neuronal excitability and sculpt neuronal growth and connections. The evident importance of these cells to chronic pain is probably a specific example of a general principle of long-term change in the nervous system. Indeed, ideas on neuronal change that fail to include glial interactions may be doomed to be incomplete¹⁰.

A three-step picture of the role of glia in chronic pain is now emerging. First, damage to, or inflammation of, a peripheral nerve is somehow communicated to microglia in the spinal cord. Second, these cells congregate, are activated, and respond to ATP by using newly produced P2X₄ receptors. And third, the ATP-activated glial cells modify signalling between spinal neurons.

This picture raises many questions, and answers to any of them will suggest new ways of relieving pain. How do spinal microglia 'know' that a nerve is damaged in the periphery? Where does the extracellular ATP come from? What signalling pathway within microglia is triggered by P2X₄ receptors? How do the microglia communicate with and alter spinal neurons? And the two most immediate questions: is the likelihood of chronic pain lessened in humans if microglial activation is prevented following nerve

damage? And is selective inhibition of P2X₄ receptors a means of doing this? Some of these questions are the domain of pain neurobiologists, but others will require immunologists and glial experts to join a field that has a goal — decreasing pain — as old as suffering itself. ■

Edwin W. McCleskey is at the Vollum Institute, Oregon Health and Science University, Portland, Oregon 97239, USA.
e-mail: mccleske@ohsu.edu

1. Dworkin, R. H. *Clin. J. Pain* **18**, 343–349 (2002).
2. Woolf, C. J. & Salter, M. W. *Science* **288**, 1765–1769 (2000).
3. Watkins, L. R. & Maier, S. F. *Physiol. Rev.* **82**, 981–1011 (2002).
4. Tsuda, M. *et al.* *Nature* **424**, 778–783 (2003).
5. Melzack, R. & Wall, P. D. *Science* **150**, 971–979 (1965).
6. Kim, S. H. & Chung, J. M. *Pain* **50**, 355–363 (1992).
7. North, R. A. *Physiol. Rev.* **82**, 1013–1067 (2002).
8. Milligan, E. D. *et al.* *J. Neurosci.* **23**, 1026–1040 (2003).
9. Nelson, P. T., Soma, L. A. & Lavi, E. *Ann. Med.* **34**, 491–500 (2002).
10. Haydon, P. G. *Nature Rev. Neurosci.* **2**, 185–193 (2001).

Solid-state physics

Drawing quantum circuitry

Ray Ashoori

An atomic-force microscope, operating at low temperatures, can be used to create and tailor quantum electronic devices, and could provide a means of building the intricate circuits needed for quantum computing.

It is 30 years since the appearance of the first semiconductor electronic devices built to make use of the effects of controllable quantum confinement of electrons. These 'resonant tunnelling diodes'¹ trapped charges in only one dimension. Since then, researchers have learned to use electrostatic gating techniques to create a variety of structures that can trap electrons in all three spatial dimensions — such as point contacts (constrictions connecting two reservoirs of electrons) and quantum dots (isolated regions of trapped electrons) — within a larger semiconductor. The simplest structure, the point contact, can function as a powerful sensor of electric charge, capable of single-electron sensitivity²; and quantum dots could serve as qubits (quantum bits) for quantum computing applications³.

Through careful lithography, experimenters have created simple circuits containing a few quantum circuit elements^{4,5}. But producing large circuits with many quantum devices has proved difficult, largely because tiny differences in fabrication result in devices with substantially different characteristics. Moreover, even with atomically precise fabrication, small amounts of static charge can accumulate randomly on surfaces, creating uncontrollable offsets in turn-on voltages and altering the device characteristics. On page 751 of this issue, Crook *et al.*⁶ demonstrate a lithographic method for producing quantum circuit elements that has

the potential to defeat both the fabrication-related problem of device variation and the problem of stray charges. It is also sufficiently flexible to permit the production of a wide variety of quantum circuits.

There are already various methods for the nanofabrication of uniform structures. For example, chemists can produce large numbers of nearly identical nanoscale solid particles⁷; through quantum confinement effects, different-size nanoparticles fluoresce with different colours, and this has been exploited in numerous applications in optics. But future quantum electronics will require more precise and well-defined connections between components than are necessary in optical circuits. Reliable techniques for such an intricate assembly of particles do not yet exist.

Scanning probe techniques might provide a way forwards. For instance, the scanning tunnelling microscope (STM) permits the controlled placement of single atoms to confine electrons in atomically precise cages on the surface of a metal⁸. This approach is different from that used to create circuits in modern semiconductor chips and in most existing quantum circuit elements, which consist nearly entirely of device elements located underneath the (insulating) surface of a semiconductor, with wiring and control gates on top. The surface control gates permit large changes in the electron density in the buried components below, allowing

substantial variation of current and coupling between different regions of the semiconductor — capabilities that are currently unfeasible in these STM-created structures. This type of gating also figures centrally in schemes for quantum computing^{3,9}. Ideally, lithographic methods for quantum devices should be based on existing schemes for producing integrated circuits with buried components.

The experiments performed by Crook *et al.*⁶ exploit the power of the atomic force microscope (AFM) to create features on an insulating surface. Crook *et al.* start with samples constructed from gallium arsenide wafers (very similar to those often used to produce high-speed transistors and also quantum components such as quantum dots and point contacts). In these structures, a bias voltage applied to metal gates patterned on the surface selectively enhances or depletes the electron density in a thin lower layer of aluminium gallium arsenide (effectively a two-dimensional 'gas' of electrons), buried beneath an insulating layer 100 nm thick that forms the sample surface. Crook *et al.* similarly prepared two metal gates, several micrometres apart, on the surface of their gallium arsenide sample (Fig. 1). With the application of a bias voltage, electron depletion left only a narrow channel — about 1 µm wide and 5 µm long — in the buried layer, connecting two large regions of electron gas.

Then, at very low temperature (about 20 mK), Crook *et al.* scanned the metallic tip of an AFM just above the sample surface, on top of or near this channel, or 'quantum wire'. By varying the voltage on the tip, they could deposit electric charge directly onto the sample surface. Negative tip voltages deposit negative charge, and this charge remains fixed indefinitely on the sample surface, repelling electrons in the layer below. Electrons could thus be confined in regions within the buried layer by simply drawing the desired structure on the top surface of the semiconductor. But the deposited charge can also be neutralized if the tip voltage is turned from negative to positive. The method has similarities to a previous scheme for AFM lithography that permanently altered samples and did not permit erasure¹⁰. This new flexibility provides a profound capability: researchers can now simply draw nanoscale electronic structures with an AFM tip, and then make adjustments or erase them completely using the same tip.

Named 'erasable electrostatic lithography' (EEL), this technique also makes it possible to measure and test a device *in situ*, as it is created — a much faster process than using conventional lithography. The structure drawn in the buried layer can be tuned to have the properties desired, and multiple structures can be linked with precisely adjusted connection strengths. Quantum engineers could develop complex circuits by

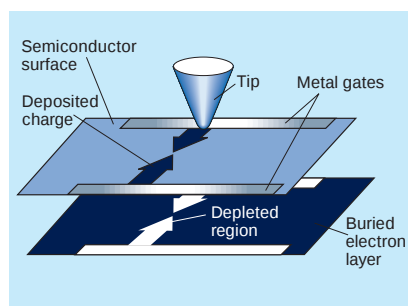


Figure 1 Erasable electrostatic lithography. A bias voltage applied to metal gates on the surface of a gallium arsenide wafer influences the electron density in an aluminium gallium arsenide layer, 100 nm below. Electron depletion creates a 'quantum wire' between the plates. Crook *et al.*⁶ have shown that moving the tip of an atomic force microscope at negative voltage across the surface deposits a negative charge on the surface, further altering the electron distribution below. The authors were able to 'draw' a constriction on the surface (blue-black areas), creating a quantum component known as a point contact in the buried layer (white areas).

first patterning a system of metal gates and then filling in the details of the circuits with the scanning probe.

In its present form, EEL works only at very low temperatures. But at the moment this is not really a drawback, because the quantum components created by EEL currently function only at low temperatures. Furthermore, if insulators were deposited on the sample surface before EEL patterning, they might trap charge better for higher-

temperature operation. A major limitation of EEL is, of course, that it requires serial drawing and tuning of each element in the circuit. This contrasts sharply with lithography used for computer chips, in which hundreds of millions of circuit elements are defined by a projection technique with only a few exposures. However, significant quantum computation could proceed with just a few thousand qubits.

Aside from engineering circuits, the use of EEL might also lead to a better understanding of the physics of quantum devices. For instance, a long-standing mystery exists in certain features of the conductance of point contacts that some theorists attribute to the shape of the constriction. In contrast with traditional lithography, by using EEL this shape could be easily varied and the effects monitored. Physicists can now literally sketch quantum devices and circuits from imagination into reality.

Ray Ashoori is in the Department of Physics, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139-4307, USA.
e-mail: ashoori@mit.edu

1. Chang, L. L., Esaki, L. & Tsu, R. *Appl. Phys. Lett.* **24**, 593–595 (1974).
2. Field, M. *et al. Phys. Rev. Lett.* **70**, 1311–1314 (1993).
3. Loss, D. & DiVincenzo, D. P. *Phys. Rev. A* **57**, 120–126 (1998).
4. Buks, E., Schuster, R., Heiblum, M., Mahalu, D. & Umansky, V. *Nature* **391**, 871–874 (1998).
5. Potok, R. M. *et al. Phys. Rev. Lett.* **91**, 016802 (2003).
6. Crook, R. *et al. Nature* **424**, 751–754 (2003).
7. Murray, C. B., Norris, D. J. & Bawendi, M. G. *J. Am. Chem. Soc.* **115**, 8706–8715 (1993).
8. Crommie, M. F., Lutz, C. P. & Eigler, D. M. *Science* **262**, 218–220 (1993).
9. Kane, B. E. *Nature* **393**, 133–137 (1998).
10. Held, R. *et al. Appl. Phys. Lett.* **73**, 262–268 (1998).

Global change

The heat on Lake Tanganyika

Dirk Verschuren

Warming of surface waters and declining fish catches in Lake Tanganyika have been linked to global climate change. The impact of global warming on natural ecosystems may be starting to affect local economies.

The effects of global climate change on ecosystems and the geographical distribution of species are already clearly visible^{1,2}, but concrete examples of their impact on the livelihood of sizeable human populations are still scarce. Now two groups — Verburg *et al.* writing in *Science*³ and O'Reilly *et al.* on page 766 of this issue⁴ — have combined documentary and field data on long-term ecosystem dynamics in Lake Tanganyika, East Africa, to show how increased heat accumulation by this deep tropical lake, linked to climate warming, has led to a marked reduction in fish yields.

Lake Tanganyika (Fig. 1, overleaf) is one of the world's great freshwater ecosystems. It is the second deepest lake and second largest

by volume (after Lake Baikal), the second largest tropical lake by surface area (after Lake Victoria), and holds the second greatest biological diversity (after Lake Baikal). This great species richness is primarily accounted for by endemic fishes, snails and crustaceans that inhabit nearshore environments. Its high ecosystem productivity, however, with a carbon transfer from algae to fish that is comparable to that of the most efficient marine fisheries⁵, comes mostly from the offshore, open-water food web, which is relatively poor in species.

At 60% by weight of overall catch, sardines are the dominant commercial fish in Lake Tanganyika. But since the late 1970s, the annual sardine catch has declined by

30–50%, to 165,000–200,000 tonnes. Since overexploitation is at most a local problem on some fishing grounds, the principal cause of this decline has remained unknown. Other human activities within the lake's drainage basin, most notably the rapidly intensifying agriculture, would be expected to increase rather than reduce aquatic ecosystem productivity, through stimulation of algal growth by runoff of extra nutrients.

Verburg *et al.*³ and O'Reilly *et al.*⁴ now propose that global climate change is responsible. They argue that rising air and surface-water temperatures over the past century have increased the thermal density contrast between surface and deep water in Lake Tanganyika. Assisted by rather slack winds in recent decades, this stronger density contrast has reduced the effectiveness of wind-driven mixing in bringing nutrients essential for primary production from the deep-water nutrient reservoir to the surface. This has resulted in decreasing algal biomass³ and consequent starvation of the upper levels of the aquatic food web, including the economically important zooplankton-eating fishes⁴.

The sensitivity of Lake Tanganyika's productivity to reduced mixing is due to its tropical location and extreme depth (1,470 metres). In temperate freshwater lakes, seasonal cooling of the lake surface breaks down the thermal density contrast between surface and deep water, and the complete mixing that results recycles the nutrients that had accumulated in the deep water after their release from decaying algae. In tropical lakes, where mixing is due to evaporative cooling of surface waters during the windy season, these recycled nutrients are crucial for maintaining lake productivity. This is because 'endless summer'⁶ conditions of high temperatures year-round accelerate all biological processes, and new nutrient inputs from the atmosphere or rock weathering cannot keep up with the high rates of algal photosynthesis and decomposition. But in the deep and steep-sided Lake Tanganyika, seasonal mixing affects only the upper reaches of the deep-water nutrient reservoir, and most nutrients entering the lake are lost from the ecosystem after a single cycle through the food web. As nutrient recycling provides 95% of the phosphorus and 97% of the silicon used by Lake Tanganyika algae in primary production⁷, modest changes in mixing efficiency have great consequences for lake productivity.

The two new studies are complementary, in that their combined evidence covers all of the important links in the chain of cause and effect between climate warming and the declining fishery. Some observations are particularly compelling. For example, little change in thermal density contrast and water transparency (an indicator of algal



Figure 1 Lake Tanganyika, seen from the Burundi highlands. Inset, sardines, the main bounty of the lake. Since the late 1970s, however, the annual catch has declined by 30–50%.

abundance) occurred between the mid-1940s and mid-1970s³, when local mean annual air temperature was relatively stable⁴. Also, whereas there used to be large seasonal variation in sardine catches, driven by the sardine cohort benefiting from algal and zooplankton blooms following the period of nutrient upwelling⁷, seasonal variation in catch is now muted⁸.

Verburg *et al.*³ use profiles of temperature change with depth, collected since 1913, to show that thermal density gradients between 110 and 200 m, and between 200 and 800 m, have tripled over the past century. The exact size of this change is clearly dependent on the depth intervals chosen, but that the density contrast between surface and deep water has increased markedly is supported by the doubling of overall (0–700 m) water-column stability calculated by O'Reilly *et al.*⁴.

O'Reilly *et al.* also present sediment records showing that the carbon-isotope composition of buried algal organic matter has become lighter over the past half-century. The authors attribute this to a decline in primary production consistent with the observed reduction in fish yields. This interpretation must assume that algae primarily benefit from upwelling of nitrogen and phosphorus present at 50–100 m depth, and that deeper mixing, which would also bring isotopically light dissolved carbon to the surface, is limited. This is somewhat uncertain, however; the case would benefit from an annual isotopic mass-balance analysis that includes algal carbon-isotopic and production data for seasons when the water column is mixed and when it is stratified. Taken together, however, the data in the two papers provide strong evidence that the effect of global climate change on regional temperature has had a greater impact on Lake Tanganyika than have local human activities.

For developing economies in Africa, the importance of the resources and services provided by the continent's great lakes and rivers can hardly be overstated. Like the other

great East African lakes (Victoria, Malawi, Kivu, Edward, Albert and Turkana), Lake Tanganyika's ecosystem is under increasing pressure from human activities in its drainage basin. Thanks to the incomplete mixing of its deep waters, surface-water quality has been buffered against the effects of excess nutrient input from deforestation and agriculture. So far, the biological consequences seem to be limited to local effects of increased sediment from eroding agricultural lands, which tends to obliterate substrate heterogeneity in nearshore bottom habitat⁹. This contrasts with the shallower Lake Victoria, where efficient nutrient recycling during annual mixing has produced a rapid biological response to excess nutrient input¹⁰. The result has been runaway algal growth, a transition to dominance of algal groups that are less efficiently used by higher levels of the aquatic food web, and a serious reduction in water quality, even though Victoria is the largest African lake by surface area.

The seven great lakes of East Africa are all dissected by international boundaries. Lake

Tanganyika is shared by the countries of Burundi, Tanzania, Zambia and the Democratic Republic of Congo, and so systems for sustainable management and conservation are subject to the difficulty of reaching multilateral agreements in a politically unstable region. Creating such systems is also hampered by our fragmentary understanding of the lakes' functioning. At present it is probably enough to provide a qualitative explanation for the distinct responses of lakes Tanganyika and Victoria to the effects of human activities. But to quantify the exact causes, evaluate the consequences and find preventative measures, much more data are needed, as are realistic mass-balance models based on a better knowledge of the inputs, outputs and biogeochemical cycling of essential nutrients in the water column.

In the meantime, we must hope that managers do not mistakenly conclude from these studies that fish yields in Lake Tanganyika can be restored by stimulating the fertilization brought by deforestation and agriculture, or that the precarious condition of Lake Victoria may be cured over time by the expected effects of global warming on water-column stability and nutrient cycling. ■

Dirk Verschuren is in the Department of Biology, Ghent University, Ledeganckstraat 35, B-9000 Gent, Belgium.

e-mail: dirk.verschuren@UGent.be

1. Walter, G.-R. *et al.* *Nature* **416**, 389–395 (2002).
2. Parmesan, C. & Yohe, G. *Nature* **421**, 37–42 (2003).
3. Verburg, P., Hecky, R. E. & Kling, H. *Science* **301**, 505–507 (2003).
4. O'Reilly, C. M., Alin, S. R., Plisnier, P.-D., Cohen, A. S. & McKee, B. A. *Nature* **424**, 766–768 (2003).
5. Hecky, R. E., Fee, E. J., Kling, H. J. & Rudd, J. W. *Trans. Am. Fish. Soc.* **110**, 336–345 (1981).
6. Kilham, P. & Kilham, S. S. *Freshwat. Biol.* **23**, 379–389 (1989).
7. Coulter, G. W. (ed.) *Lake Tanganyika and its Life* (Oxford Univ. Press, 1991).
8. van Zwieten, P. A. M., Roest, F. C., Machiels, M. A. M. & van Densen, W. L. T. *Fish. Res.* **54**, 329–348 (2002).
9. Alin, S. R. *et al.* *Conserv. Biol.* **13**, 1017–1033 (1999).
10. Verschuren, D. *et al.* *Proc. R. Soc. Lond. B* **269**, 289–294 (2002).

Cancer

Protective packaging for DNA

Jessica A. Downs and Stephen P. Jackson

Histone proteins are best known for their structural role in packaging DNA into a compact form. But it seems that one such protein also works as a tumour suppressor, helping to prevent cancer developing.

Inside every human cell there is an impressive two metres or so of DNA, which must be organized into a compact form to fit into the cell nucleus. To do this, cells begin by wrapping the DNA around histone proteins. But one of these proteins, histone H2AX, has a dual role — it also helps cells to deal with potentially dangerous breaks that can occur in the DNA double helix. Writing last week in *Cell*, Celeste *et al.*¹ and

Bassing *et al.*² highlighted the importance of this function: they showed that it helps to prevent genome instability and cancer.

The compact form of DNA in the nucleus is known as chromatin. The basic unit of chromatin, the nucleosome, is composed of DNA wrapped around two copies each of histone proteins H2A, H2B, H3 and H4. The H2AX variant makes up about 10–15% of total cellular H2A in higher organisms

such as humans, frogs and mice, and is incorporated randomly into nucleosomes throughout the genome. It differs from H2A primarily in that it has an extended tail at one end — the carboxy terminus. Following DNA damage, this tail rapidly becomes labelled with phosphate groups, generating a species called γ -H2AX³. In lower organisms such as yeast, the carboxy-terminal tail of H2A itself contains a similar motif to the H2AX tail, and becomes phosphorylated when DNA is damaged.

Although it is unclear exactly what γ -H2AX does following DNA damage, microscopy studies have shown that it is generated in the chromatin flanking a DNA double-strand break, and that mammalian repair and signalling proteins are recruited to these sites, ultimately amassing in large numbers. These visible protein accumulations, which can span millions of bases of DNA, are known as foci. γ -H2AX is not required for the initial recruitment of repair factors, but is needed for later foci formation⁴.

Wherever possible, the cell's aim is to repair the DNA break, which can be done in various ways — two of the most common methods are non-homologous end-joining (NHEJ) and homologous recombination. Previous work² indicates that mutating the phosphorylation site on the carboxy terminus of yeast H2A impairs, but does not prevent, the repair of double-strand breaks by NHEJ. No obvious NHEJ defect has been observed in mouse cells lacking H2AX, although their mechanism for repairing double-stranded breaks by homologous recombination seems to be defective^{6,7}. H2AX is also required following low (but not high) levels of DNA damage in order to trigger the 'checkpoint' that halts cell division to allow time for DNA repair⁸. The new papers^{1,2} help us to integrate these seemingly disparate findings into a more coherent model (Fig. 1).

Because defects in the detection and repair of double-strand breaks can cause cancer, Celeste *et al.*¹ and Bassing *et al.*² wanted to see whether H2AX functions as a tumour suppressor — a protein that hampers tumour development. To do this, they generated mice in which either one or both copies of the H2AX gene were inactivated, either alone or in combination with inactivation of the tumour suppressor p53. Strikingly, although mutations in H2AX alone engendered little or no increase in cancer development, they dramatically enhanced tumour formation when combined with p53 deficiency.

The authors also found evidence of genome instability in the mice that lacked both H2AX and p53. Analysis of cancers arising from B cells, a type of immune cell, revealed that segments of DNA had been moved (translocated) from one chromosome

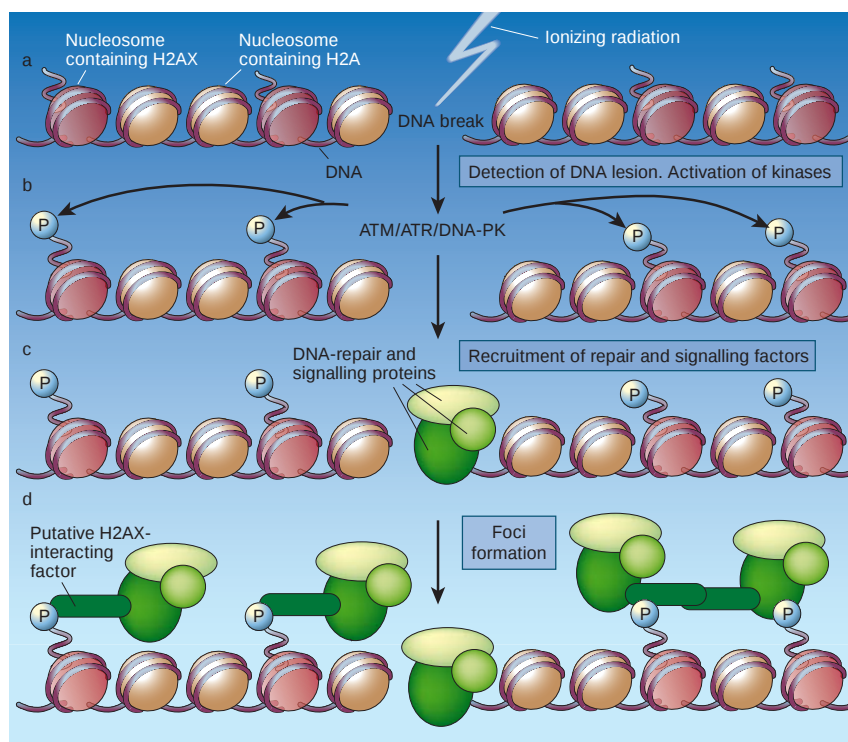


Figure 1 Model of the response to DNA double-strand breaks in mammalian cells, incorporating the new results^{1,2}. a, Double-strand breaks are produced by, for instance, ionizing radiation, or during the rearrangement of immune-receptor genes. b, The lesion is detected and various kinase enzymes (ATM, ATR and DNA-PK) are activated, in turn phosphorylating histone H2AX protein. ATM, ATR and DNA-PK also transduce the signal to downstream proteins (not shown), leading to the activation of checkpoints that prevent cell division when levels of DNA damage are high, and the production of various repair and signalling factors, which are recruited to the lesion (c). d, Further factors are recruited, aided by the phosphorylated H2AX, and perhaps involving intermediary proteins. H2AX may therefore enhance accurate, efficient repair, or prevent inaccurate or inefficient repair. H2AX phosphorylation may also lead to changes in chromatin, or facilitate a checkpoint that prevents cell division when there is little DNA damage.

to another. The regions involved indicated that the translocations were failed rearrangements of immunoglobulin genes. These types of rearrangements generate receptor diversity among both B and T cells in the immune system, and they occur through an intermediate formed by a double-strand break that is, at least in some cases, repaired by NHEJ. Such translocations were, however, rarely found in T-cell malignancies, but these frequently harboured rearrangements indicative of inappropriate joining of spontaneously arising double-strand breaks. Significantly, chromosomal breaks and translocations, some possibly involving the immune-receptor genes, are also seen in mice that lack H2AX but have p53 (ref. 7). The lack of marked tumour formation in these circumstances presumably indicates that cells undergoing cancer-causing genome rearrangements are eliminated by p53-dependent cell death.

Some of the effects of H2AX deficiency^{1,2} are similar to those seen in mice deficient in NHEJ, in that these — when also deficient in p53 — develop cancers that are associated with frequent translocations involving

immune-receptor genes^{9,10}. Taken together, then, it seems that the fidelity of NHEJ is partially disrupted when H2AX is impaired. This is consistent with the recent report that some immunoglobulin rearrangements are defective in H2AX-deficient mice even when p53 is present¹¹.

So H2AX joins the growing number of genome 'caretakers' that also function as tumour suppressors. Notably, Celeste *et al.* and Bassing *et al.* show that p53-deficient mice with one defective copy of the H2AX gene also develop tumours, albeit at lower rates than their wholly H2AX-deficient counterparts. Furthermore, the normal copy of H2AX is retained in these tumours. In genetic parlance, then, H2AX displays 'haploinsufficiency' — it functions in a dose-dependent manner. And, unlike the situation with 'classical' tumour suppressors, complete loss of H2AX function is not required for tumour development.

These are exciting findings, but the question remains: how does the phosphorylated form of H2AX influence the repair of double-strand breaks and hence genome stability? As repair still occurs in the absence

of H2AX, it is improbable that this protein is required in the catalytic steps of the process. As highlighted by the new reports^{1,2}, it seems instead to affect the efficiency or accuracy of the repair.

One clue to how it might do this comes from the fact that foci formation is impaired in its absence. Perhaps the phosphorylation of the H2AX tail produces a binding site for a protein or complex that facilitates the accumulation of checkpoint and repair proteins at sites of damage, following the initial recruitment of a few such proteins. This accumulation of factors might stabilize the DNA ends, perhaps maintaining them in proximity and so preventing inappropriate repair. Also, foci formation may be the mechanism by which H2AX mediates the cell-cycle checkpoint when levels of DNA damage are low. Other DNA-damage-induced covalent modifications, either on the H2AX tail or elsewhere in the nucleosome, might influence the interactions of γ -H2AX.

Another, not necessarily exclusive, possibility is that phosphorylation of H2AX directly affects chromatin structure. The H2AX tail is not a part of the nucleosome core, but extrudes outwards, in the region where the DNA enters the nucleosome. Phosphorylation here could easily be envisaged to affect the association of DNA with a nucleosome, the binding of linker histones, or the ability of nucleosomes to fold into higher-order structures. This could, in turn, provide a more amenable template for manipulation by repair factors, or it could stabilize the DNA ends as above. There is *in vivo* evidence to support this hypothesis^{3,12}, although it is not known whether chromatin structure is being directly or indirectly altered by histone phosphorylation.

Whatever the mechanism, Celeste *et al.*¹ and Bassing *et al.*² have shown that a structural component of chromatin can also act as a tumour suppressor in mice. The same might be true in humans, as the H2AX gene maps to a region of the genome that is frequently mutated in human tumours. If so, this could have medical applications. For instance, the H2AX 'status' of a person might help to define their underlying predisposition to cancer. Moreover, a look at p53 and H2AX activity in a particular tumour could ultimately help clinicians in their choice of therapy.

Jessica A. Downs is in the Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge CB2 1GA, UK.

e-mail: js4@mole.bio.cam.ac.uk

Stephen P. Jackson is at the Wellcome Trust/Cancer Research UK Institute, Tennis Court Road, Cambridge CB2 1QR, UK.

e-mail: spj13@mole.bio.cam.ac.uk

1. Celeste, A. *et al.* *Cell* **114**, 371–383 (2003).

2. Bassing, C. H. *et al.* *Cell* **114**, 359–370 (2003).

3. Redon, C. *et al.* *Curr. Opin. Genet. Dev.* **12**, 162–169 (2002).

4. Celeste, A. *et al.* *Nature Cell Biol.* **5**, 675–679 (2003).

5. Downs, J. A., Lowndes, N. F. & Jackson, S. P. *Nature* **408**, 1001–1004 (2000).

6. Bassing, C. H. *et al.* *Proc. Natl Acad. Sci. USA* **99**, 8173–8178 (2002).

7. Celeste, A. *et al.* *Science* **296**, 922–927 (2002).

8. Fernandez-Capetillo, O. *et al.* *Nature Cell Biol.* **4**, 993–997 (2002).

9. Difilippantonio, M. *et al.* *Nature* **404**, 510–514 (2000).

10. Gao, Y. *et al.* *Nature* **404**, 897–900 (2000).

11. Reina-San-Martin, B. *et al.* *J. Exp. Med.* **197**, 1767–1778 (2003).

12. Fernandez-Capetillo, O. *et al.* *Dev. Cell* **4**, 497–508 (2003).

Atmospheric physics

A new dawn for aurora

Patrick T. Newell

A new type of aurora has been observed. Although similar in shape to conventional aurora, it lacks their intensity and dynamics, instead rotating quietly with the Earth.

'Aurora' is the term given to the glow created when energetic particles from space strike Earth's upper atmosphere, or ionosphere. Aurora — also known as the northern and southern lights (*aurora borealis* and *aurora australis*, respectively) — usually occur within rings around the magnetic poles. Each ring is up to about 3,500 km in diameter and a few hundred kilometres deep. This 'auroral oval' contains discrete aurora, which form curtain or ray shapes that are often easily visible with the naked eye (Fig. 1), and diffuse aurora, which are fainter and lack sharp spatial boundaries. In *Geophysical Research Letters*, Kubota *et al.*¹ report new auroral observations that fit neither category and lie several degrees below the auroral oval. Most intriguingly, the sub-oval aurora appears to co-rotate with the Earth. These observations suggest that an old idea² about how aurora get their shapes may yet prove to be true.

Kubota *et al.*¹ used high sensitivity all-sky cameras at Poker Flat, 30 miles north of Fairbanks, Alaska. Much of the time, the solar wind drives a flow of ionized particles, or

plasma, around the Earth; as a result, electric current flows from space into the upper atmosphere, creating moderate or high geomagnetic activity and bright aurora. Under such conditions, the Poker Flat observing station lies within the auroral oval, and fulfils its intended function of studying discrete and diffuse auroral forms. When geomagnetic conditions are quiet, however, the position of the auroral oval lies polewards of Poker Flat, and no aurora are expected at this site. Yet under these conditions, Kubota *et al.* discovered faint auroral structures, with a distinctly different phenomenology (Fig. 1). These sub-oval aurora have the ribbon-like appearance typical of discrete aurora within the auroral oval. But they also have the quiescence, persistence and low intensity that are more typical of diffuse aurora. Data from an overflying US Air Force satellite showed no acceleration of the electrons impacting on the atmosphere — as is the case for conventional discrete aurora.

The sub-oval aurora also move with the Earth, at about 70% of its velocity, drifting at 140 km s⁻¹ in the direction of Earth's spin. This is a surprise, and a suggestive one. Aurora tend to be positioned such that an observer

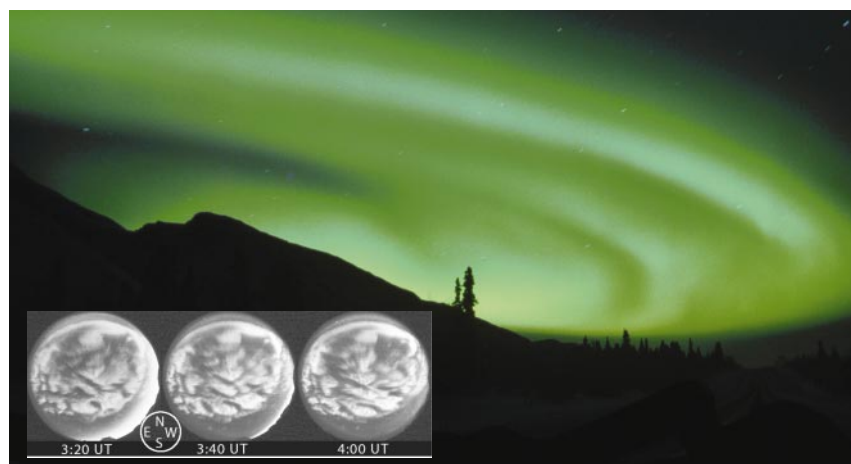


Figure 1 Northern lights. Bands of green light across the Alaskan sky form the *aurora borealis*, usually seen inside the auroral oval that surrounds the Earth's magnetic north pole. But Kubota *et al.*¹ have detected a new type of aurora (inset) that occurs at latitudes below the auroral oval and is phenomenologically quite different.

P. A. SOUTHERS/CORBIS

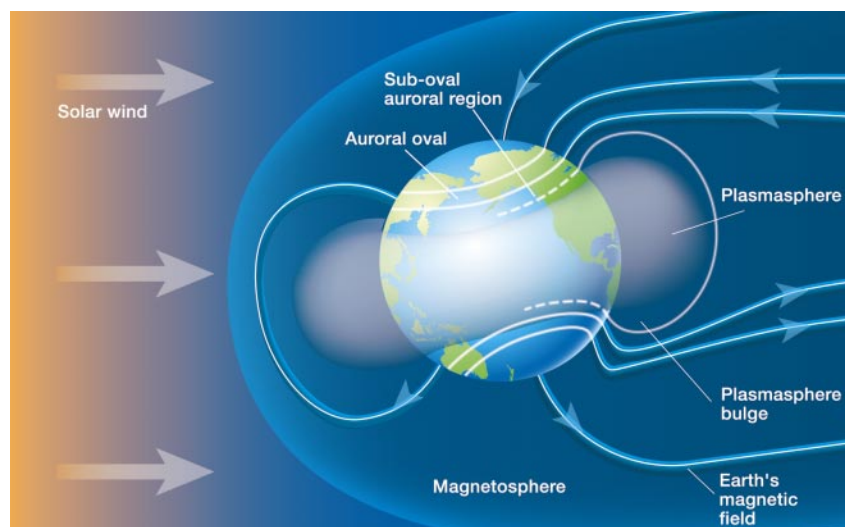


Figure 2 The Earth's magnetic environment. Our planet's magnetic field forms a protective bubble (the magnetosphere) that excludes most of the solar wind. The plasmasphere — a region of cold plasma — fills near-Earth space, and bulges outwards over the part of the planet that is at dusk. Energetic particles from space striking the atmosphere create aurora, which are generally contained within the auroral ovals around the planet's magnetic poles. But Kubota *et al.*¹ have observed a new type of aurora over Alaska, at lower latitude than the auroral oval. The new sub-oval aurora are quite different to those usually observed in the auroral oval, and may be connected to the impingement of the plasmasphere's bulging boundary on the atmosphere below.

on Earth sees auroral patterns apparently drifting westwards, as Earth's spin actually rotates him eastwards. Kubota and colleagues' faint co-rotating aurora must therefore be more closely linked to the near-Earth region than are previously known aurora.

A few earlier, related observations do exist. Photographs from the ISIS-2 satellite³ showed instances of 'detached aurora', lying below the main auroral oval. Only one image of these aurora was available for each orbit of the satellite, and little could be learned from them. However, there has been much discussion^{4,5} in the past few years of potentially related phenomena, such as sub-auroral ion drifts and sub-auroral polarization electric fields. Interest in these phenomena has been sparked by clear evidence⁶ that the upper atmosphere at middle latitudes has a surprisingly strong connection to activity in the auroral oval and to the solar wind. Kubota and colleagues' work seems to show that these are not, in fact, sub-auroral phenomena, despite lying below the established oval of discrete and diffuse aurora. Whether they are related in some way to the new sub-oval aurora is not yet clear.

More than two decades ago, Carl McIlwain (co-discoverer of the Van Allen radiation belts that surround Earth, with James Van Allen) proposed an explanation² for why some aurora appear latitudinally narrow but longitudinally elongated. Earth is surrounded by a plasmasphere — a co-rotating sphere of cold plasma, which flows up from the ionosphere (Fig. 2). Over the sector of Earth experiencing dusk, the plasmasphere often bulges outwards, in

response to the Earth's rotation and to the electric fields induced by the solar wind. McIlwain thought that the aurora might be shaped by the protrusion, following Earth's magnetic-field lines, of the swelling boundary of the cold plasma into the ionosphere below. But his idea proved not to fit the aurora observed within the auroral oval — diffuse aurora are too spatially unstructured to match a boundary, and discrete aurora have rapid dynamic motions resulting from lower-altitude coupling of the ionosphere and the magnetosphere.

The quiescent, discrete, elongated aurora discovered by Kubota *et al.*¹, however, fit the bill. Their near co-rotation with the Earth also supports McIlwain's model, which Kubota *et al.* seem to have independently resurrected. If these findings and associations are confirmed, they could help to explain the unexpectedly strong connection between the solar wind, the aurora and the composition and electron density of Earth's upper atmosphere, even at latitudes that are nominally below the auroral oval. ■

Patrick T. Newell is in the Applied Physics Laboratory, Johns Hopkins University, Laurel, Maryland 20723, USA.
e-mail: patrick.newell@jhuapl.edu

- Kubota, M., Nagatsuma, T. & Murayama, Y. *Geophys. Res. Lett.* doi:10.1029/2002GL016652 (2003).
- McIlwain, C. E. in *Physics of Auroral Arc Formation* (eds Akasofu, S.-I. & Kan, J. R.) 173–174 (Am. Geophys. Union, Washington DC, 1981).
- Wallis, D. D. *et al. J. Geophys. Res.* **84**, 1347–1360 (1979).
- Anderson, P. C. *et al. J. Geophys. Res.* **106**, 29585–29599 (2001).
- Foster, J. C. & Vo, H. B. *J. Geophys. Res.* doi:10.1029/2002JA009409 (2002).
- Craven, J. D. *et al. Geophys. Res. Lett.* **21**, 2793–2796 (1994).



100 YEARS AGO

A Gloucestershire Wild Garden. By the Curator (London: Elliot Stock, 1903.)
Price 6s. net.

Gardening books are becoming noted for containing a small amount of gardening information largely diluted with something that has little or no relevance to horticultural pursuits. The diluting medium may be cookery or hygiene, tirades against vivisection, stale jokes, spiritualism, anything, in fact. In the present book, gardening, or one phase of it, represents the slices of bread between which are inserted, sandwich-fashion, dissertations on the molecular structure of the brain and nerve centres, and discussions on the origin of thought and the nature of religious impressions. The "Curator" is the gardener who evidently knows plants and loves them. To him appear when he is tired of work, or, at any rate, without preface or apology, a somewhat prosy "Professor," who supplies the anatomical details above mentioned, and explains them from the materialistic standpoint, and an orthodox "Padre," who is somewhat shocked at the views propounded by the professor. The Curator acts as moderator, and when discussion seems likely to become dangerous, suggests a pipe of tobacco or a cup of tea as effectual "shunters." At any rate, we pass abruptly from metaphysical subtleties either to the tea-table or to another chapter, in which we are told how to construct a "wild" garden. As if all this were not enough, a love story — a very short one — is introduced, and so the book has one quality which a garden should possess, and that is, variety.
From *Nature* 13 August 1903.

50 YEARS AGO

The role of the tricarboxylic acid cycle in the metabolism of micro-organisms has been investigated by many authors with contradictory results, and the question has been recently reviewed by Krebs. This author, working with yeast, concluded that the reactions of the tricarboxylic acid cycle are primarily concerned with the supply of metabolic intermediates rather than of energy. Experimental support for this conclusion is given by the present investigation. It is shown that citric acid, without actively participating in respiration, may be very effective in the nitrogen metabolism of a variant (M_2^-) of *B. subtilis*.
From *Nature* 15 August 1953.

Another toll road

Wen-Chen Yeh and Nien-Jung Chen

The immune response to microbes involves a well-known signalling cascade — but this doesn't control all aspects of anti-pathogen warfare. A second cascade, involving a new 'adaptor' protein, helps fill the gap.

The innate immune system enables animals to deal quickly and efficiently with invading microorganisms. A key component of this rapid-reaction force is a finely honed detection system, which consists of a set of receptors, found on the surface of immune cells, that recognize particular molecular patterns associated with pathogens. When they sense the presence of intruders, these so-called Toll-like receptors trigger an intracellular signalling pathway that culminates in various immune responses¹. This pathway has been well characterized. But it has become clear that it does not elicit all of the responses to bacteria and viruses — in particular the production of interferon- β protein². So a second signalling pathway with this specific function has been postulated. Writing on page 743 of this issue and in *Science* respectively, Hoebe *et al.*³ and Yamamoto *et al.*⁴ report that a protein variously named Trif or Ticam-1 is a key part of that second pathway. Animals with defective Trif protein, or lacking Trif altogether, are severely defective in their ability to mount inflammatory responses against viral and bacterial components.

Toll-like receptors (TLRs) come in

various subtypes, which detect different components of pathogens. When TLRs on the surface of an immune cell are engaged, two major events are triggered downstream. First, all TLRs trigger the secretion of intercellular messenger proteins called cytokines, which initiate and perpetuate inflammation. Second, the stimulation of the TLR3 and TLR4 receptors also triggers unique signals that lead to the secretion of interferon- β . This protein subsequently causes the activation of a series of genes whose products help to eliminate the pathogen. TLR3 detects the presence of double-stranded RNAs (such as some viral genomes) and similar structures, and is therefore crucial in warding off viruses⁵. TLR4 is stimulated by lipopolysaccharide, a major component of the cell wall of certain bacteria⁶.

In most cases, the production of inflammatory cytokines in response to TLR stimulation depends on a common molecular pathway that is anchored by an 'adaptor' protein called MyD88. This protein contains a structural region called a TIR domain, which allows it to associate with the intracellular tail of a TLR and to connect the receptor to molecules that propagate a signal further inside the cell. The pathway culminates in the

nucleus with the activation of appropriate genes, such as cytokine genes. But MyD88 is not involved in some responses to TLR3 and TLR4 activation, including interferon- β production². So how do such events come about?

A clue lies in the report by Hoebe and colleagues³ that mice with a chemically induced mutation called *Lps2* show severe impairment of immune responses mediated by TLR3 and TLR4, and are highly susceptible to viral infections. The *Lps2* mutation turns out to occur in the Trif protein — which has a similar receptor-interacting domain to MyD88 — and may inactivate it. The mutation thereby abolishes the activation of the gene-transcription factor IRF-3. Stimulation of this factor was known to be triggered by TLR3 and TLR4, but to be independent of MyD88; so these data distinguish the function of Trif from that of MyD88.

What happens when Trif is missing completely? Yamamoto *et al.*⁴ used conventional gene targeting to generate a Trif-deficient mouse strain, and report similar results to those of Hoebe and colleagues. Specifically, they found that Trif-deficient cells from several tissues show impaired TLR3- and TLR4-mediated responses, including all of the responses known to be independent of MyD88. In particular, the almost complete absence of TLR3-mediated signalling in Trif-null and *Lps2*-mutant cells suggests that Trif is a major adaptor protein for TLR3 signalling^{3,4}.

It is worth noting that Trif was initially identified^{7,8} as a protein that transduces signals from activated TLR3; the new findings confirm this role. What wasn't known, however, is that Trif is also essential in transducing signals from bacterial lipopolysaccharide, via TLR4, for interferon- β production (Fig. 1). Curiously, the two groups^{3,4} also show that mice with defective or missing Trif protein exhibit impaired production of inflammatory cytokines in response to lipopolysaccharide. As mentioned above, the production of such cytokines proceeds via MyD88 — so it seems that Trif can also collaborate with this protein following TLR4 activation (Fig. 1). The major signalling pathway leading to production of inflammatory cytokines involves the transcription factor NF- κ B and various mitogen-activated protein kinase enzymes. Hoebe *et al.* and Yamamoto *et al.* found that the activation of these proteins by lipopolysaccharide was only slightly impaired in mice with a mutation in either MyD88 or Trif. But there was no activation at all in mice lacking both adaptor proteins.

One advantage to the cell of using multiple adaptors and signalling pathways may be to provide response specificity: different pathogen components engage different TLRs, which recruit different combinations of adaptors, yielding different outcomes appropriate for eliminating those pathogens.

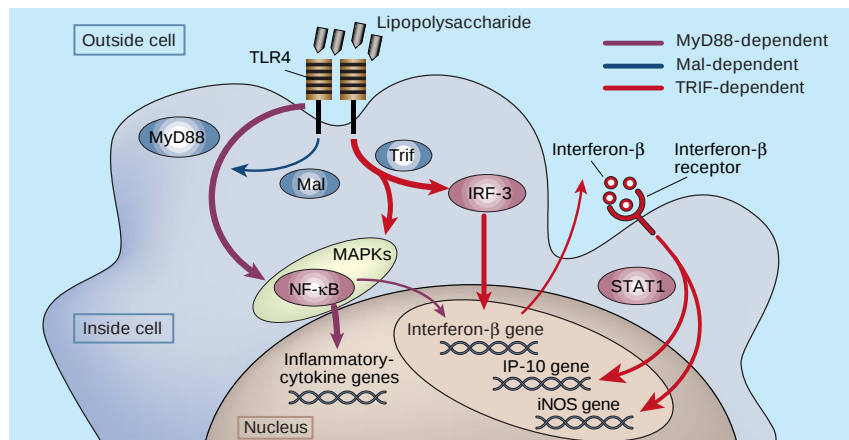


Figure 1 Fighting bacteria. A major component of the cell wall of certain bacteria is lipopolysaccharide, which is recognized by Toll-like receptor 4 (TLR4) on animals' immune cells. TLR4 then initiates the well-characterized MyD88-dependent signalling pathway, which contributes primarily to the activation of the transcription factor NF- κ B and the production of inflammatory cytokines. Hoebe *et al.*³ and Yamamoto *et al.*⁴ have identified a second pathway, mediated by a MyD88-related molecule called Trif (or Ticam-1). This pathway leads to activation of the transcription factor IRF-3, which controls the production of interferon- β . Interaction of interferon- β with its receptor initiates signalling through STAT1, which helps to activate further genes. Trif also collaborates with the MyD88 pathway by activating NF- κ B and various mitogen-activated protein kinases (MAPKs), and promoting cytokine production. Mal is another MyD88-related protein. Thicker arrows represent major pathways.

Yet another TIR-domain-containing adaptor is Mal, which collaborates with MyD88 in pathways triggered by engagement of TLR2 or TLR4 (refs 9, 10). And at least two more members of this family have been identified, although their physiological functions are unknown¹¹.

The fact that MyD88, Mal and Trif can potentially interact with TLR4 in response to lipopolysaccharide, and that all contribute to downstream signal transduction, raises several questions. Does TLR4 interact with these adaptors directly? Does it bind to them simultaneously or successively? Are the effects of the interactions synergistic, or are they mutually exclusive in a way that provides temporal regulation of different signalling pathways? Hoebe *et al.* observed that *Lps2*-mutant mice have two populations of Trif-defective macrophages (a major type of immune cell): one population is severely impaired in lipopolysaccharide-induced cytokine production whereas the other is virtually unaffected. (In contrast, all MyD88-deficient macrophages exhibit impaired lipopolysaccharide-induced cytokine production.) It remains to be seen whether the 'Trif-independent' *Lps2* macrophages naturally use an alternative adaptor in place of Trif, or whether they transduce an inherently different type of MyD88-dependent signal.

The bipartite nature of the lipopolysaccharide-induced signalling pathways that lead to cytokine production might prove useful clinically. Inflammatory responses initiated by TLRs are crucial for defence against pathogens — but they can be harmful if uncontrolled. For example, a severe bacterial infection can lead to death from septic shock, which is caused not directly by the pathogen but rather by excessive inflammation. Modulating one branch of the lipopolysaccharide-induced pathways *in vivo* might eliminate the risk of septic shock while preserving an adequate immune response. Hoebe *et al.* and Yamamoto *et al.* have shown that disrupting the Trif-dependent pathway can reduce the toxicity of lipopolysaccharide *in vivo*. It will be interesting to see how well Trif-deficient animals fare against bacteria.

One final question is which other molecules contribute to the Trif pathway: signal transduction downstream of a given molecule depends on the transducers it can recruit. Apart from certain conserved structural features in the TIR domain, the amino-acid sequence of Trif diverges sharply from that of MyD88. So Trif might recruit a different set of transducers to relay signals downstream. Identifying this molecular pathway^{12,13} will be crucial in understanding antiviral defences, and may prove useful in modulating inflammatory responses to bacteria.

Wen-Chen Yeh and Nien-Jung Chen are at the Advanced Medical Discovery Institute, University

Health Network, and the Department of Medical Biophysics, University of Toronto, 620 University Avenue, Toronto, Ontario M5G 2C1, Canada. e-mails: wyeh@uhnres.utoronto.ca nchen@uhnres.utoronto.ca

1. Janeway, C. A. Jr & Medzhitov, R. *Annu. Rev. Immunol.* **20**, 197–216 (2002).
2. Takeda, K. & Akira, S. *Cell. Microbiol.* **5**, 143–153 (2003).
3. Hoebe, K. *et al. Nature* **424**, 743–748 (2003).
4. Yamamoto, M. *et al. Science* **301**, 640–643 (2003).

5. Alexopoulou, L., Holt, A. C., Medzhitov, R. & Flavell, R. A. *Nature* **413**, 732–738 (2001).
6. Poltorak, A. *et al. Science* **282**, 2085–2088 (1998).
7. Yamamoto, M. *et al. J. Immunol.* **169**, 6668–6672 (2002).
8. Oshiumi, H. *et al. Nature Immunol.* **4**, 161–167 (2003).
9. Yamamoto, M. *et al. Nature* **420**, 324–329 (2002).
10. Horng, T., Barton, G. M., Flavell, R. A. & Medzhitov, R. *Nature* **420**, 329–333 (2002).
11. O'Neill, L. A. J., Fitzgerald, K. A. & Bowie, A. G. *Trends Immunol.* **24**, 286–289 (2003).
12. Sharma, S. *et al. Science* **300**, 1148–1151 (2003).
13. Fitzgerald, K. A. *et al. Nature Immunol.* **4**, 491–496 (2003).

Earth science

Glaciers at work

Jan A. Piotrowski

Glaciers have been powerful agents in redistributing sediment and reshaping parts of the Earth's surface. Little is known about how exactly they do it, but a new hypothesis looks set to improve matters.

Over the past two million years, glaciers have repeatedly expanded into lower latitudes, before retreating and then advancing again. Instead of the processes of erosion and deposition that have operated during most of Earth's history, and that gradually level-off landscapes, during this time large parts of the world's surface have experienced comparatively fast-acting forces that can shift large volumes of rock and soil. Older landscapes may have become completely reshaped, depending on where glaciers have removed material and then dumped it.

The results of glacier action are clear from the landscapes that are exposed as glaciers melt. But what is actually happening underneath glaciers or ice sheets, which can be hundreds of metres thick, when rocks and soils are being destroyed and displaced? On page 758 of this issue, Alley *et al.*¹ describe how they have tackled this question by combining glaciological theory and observations. They have come up with a mechanism through which sediment is redistributed by meltwater under glaciers, and which explains certain aspects of landscape evolution.

The classic chain of glacial action is as

follows. The process begins with erosion, with material being picked up from the glacier bed²; it continues with the material being transported in a debris-rich basal ice and in a 'conveyor belt' just below it, or in meltwater flowing towards the ice margin; and it ends with deposition of the sediment and larger debris when the transport mechanisms die out. Glacial erosion can vary by several orders of magnitude up to around 100 mm per year, and at these higher levels is possibly the highest sustained rate of erosion seen on Earth³. Over the past two decades, however, views on glacial erosion and transport have tended to be polarized, with emphasis on either the ice itself or its meltwater as the major player. Consensus has been conspicuous by its absence^{2–5}.

Structures called 'overdeepenings' — large basins that can be square kilometres in size — are especially spectacular examples

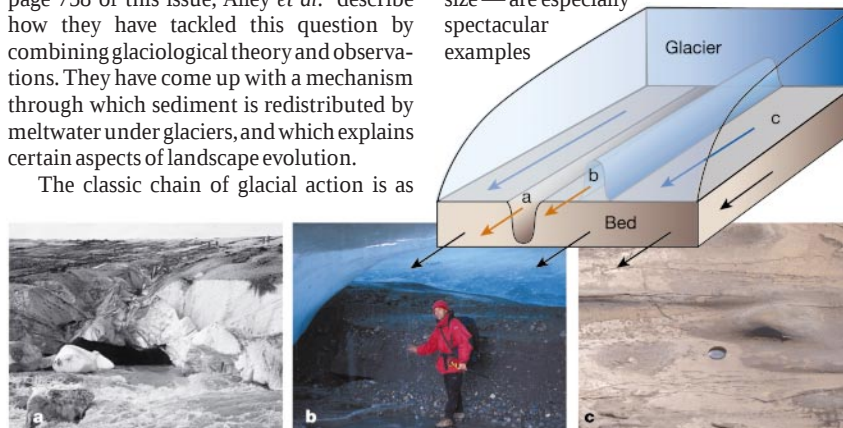


Figure 1 Three major systems of meltwater discharge from under a glacier. Discharge can occur through channels (orange arrows), in a sheet along the interface between the glacier and its bed (blue arrows) or as groundwater flow (black arrows). The photographs give examples of (a) a river emerging from a channel mouth at a glacier margin (note the figures on the glacier); (b) a channel cut into a glacier's basal ice, observed under low-water conditions; and (c), small erosional hollows left by meltwater on the bed of an old glacier (the central circular object is a lens cap). Alley *et al.*¹ now propose that a sediment pump, which is switched on and off by opening and closing meltwater channels around a threshold for water supercooling, may control a glacier's erosive forces.

of glacial erosion. They are typically found just inside the outermost extent of glaciers and have 'floors' that are seated deeper than the lowest end of the glacier. Alley and colleagues' analysis¹ of such depressions suggests that the terrain-moulding action of glaciers is organized in a predictable fashion, in which a series of stabilizing feedbacks tends to keep an equilibrium between deposition and erosion. This general principle is well established for rivers — there, for instance, more sediment will make the water shallower, which will tend to increase water flow, which in turn will tend to remove sediment.

For glaciers, Alley and colleagues' basic concept is likewise fairly simple, making their hypothesis attractive. Water formed by melting on top of a glacier may reach the bed through a system of veins, which are especially found close to the glacier edges, where the amount of melting is high and the ice is relatively thin⁶. This water, captured in channels at the bed, will then flow to the ice margin. On the way, it will cut down into the bed, removing sediment, and at the edge will excavate overdeepened basins. Erosion will be maintained by a vigorous water flow that overpowers the amount of sediment being removed.

But this process cannot go on for ever, and Alley *et al.* explain why. In channels far beneath the glacier, water does not freeze because the pressure of the overlying ice depresses the melting point — just as high pressure under ice-skates melts the ice and enables the skater to glide across the ice surface. But close to the ice margin, water channels have to ascend the adverse slope of an overdeepening before the water is expelled in front of the glacier. If the slope is sufficiently steep (about 1.5 times steeper than the slope of the glacier surface above), the falling ice pressure will cause water to freeze. This is known as glaciohydraulic supercooling, and is an idea that has been developed by work on the Matanuska Glacier in Alaska⁷ and confirmed on glaciers elsewhere⁸.

As Alley *et al.* point out, the consequences of supercooling may be dramatic: the free-flowing channels become plugged by freezing water; sediment transport capacity drops by half; and an overdeepening changes into a dumping site for sediment delivered from areas further up the glacier. If the in-fill reduces the slope of the overdeepening sufficiently, the system may switch back to water flow and erosion. This sediment pump may also be turned on and off because of fluctuations in ice thickness, and associated changes in the pressure-controlled melting point, that may allow or inhibit the existence of free water at the base of the glacier. Alley *et al.* suggest that this relationship of an overdeepening to ice thickness may make overdeepenings a sensitive proxy measure of climate change. Moreover, if the state of equilibrium between erosion and deposition is, as they postulate, close to the supercooling threshold, this

threshold may turn out to be a crucial control on glacial processes near the ice margin.

It has long been obvious that glaciers can mechanically cut, push and dump sediment, just as a bulldozer displaces earth. Such processes have been the traditional explanation for the big depressions caused by glacier action. Alley and colleagues' hypothesis brings meltwater activity (Fig. 1) to the forefront as a powerful agent that in itself can profoundly reshape terrain. If supercooling sets in, however, large amounts of material may be frozen onto the basal ice, which would be good news for those who stress the importance of sediment transport inside the ice rather than in a conveyor belt below it, and of ice motion at the ice-bed interface rather than in the underlying soft sediment^{9,10}. Such basal freezing should also deepen the basin by plucking sediment from the bed — that is, it should counteract the sedimentation caused by channel closure. But these conflicting processes cannot yet be compared quantitatively.

A next step in this line of research is to find out whether the overdeepenings seen at the edges of glaciers in modern times also developed under the terminal parts of the enormous ice sheets that were characteristic

of the ice ages. If they did, and if they fluctuated with the ice sheets in time and space, then the mechanism proposed by Alley *et al.* may provide a new tool for interpreting continental-scale patterns of sediment redistribution by ice sheets. And could meltwater processes such as those proposed by the authors also have operated far beneath the ice and well behind the leading margin? If so, they could have influenced large-scale ice-sheet dynamics, and thus a component of Earth's cold-climate control system. ■

Jan A. Piotrowski is in the Department of Earth Sciences, University of Aarhus, DK-8000 Aarhus C, Denmark.

e-mail: jan.piotrowski@geo.au.dk

1. Alley, R. B., Lawson, D. E., Larson, G. J., Evenson, E. B. & Baker, G. S. *Nature* 424, 758–760 (2003).
2. Knight, P. G. *Quat. Sci. Rev.* 16, 975–993 (1997).
3. Hallet, B., Hunter, L. & Bogen, J. *Glob. Planet. Change* 12, 213–235 (1996).
4. Boulton, G. S. *J. Glaciol.* 42, 43–62 (1996).
5. Fisher, T. G., Clague, J. J. & Teller, J. T. *Quat. Int.* 90, 1–4 (2002).
6. Fountain, A. G. & Walder, J. S. *Rev. Geophys.* 36, 299–328 (1998).
7. Alley, R. B., Lawson, D. E., Evenson, E. B., Strasser, J. C. & Larson, G. J. *J. Glaciol.* 44, 563–569 (1998).
8. Alley, R. B., Lawson, D. E., Evenson, E. B. & Larson, G. J. *Ann. Glaciol.* (in the press).
9. Iverson, N. R., Hanson, B., Hooke, R. LeB. & Jansson, P. *Science* 267, 80–81 (1995).
10. Piotrowski, J. A., Mickelson, D. M., Tulaczyk, S., Krzyszkowski, D. & Junge, F. *Quat. Int.* 86, 139–150 (2001).

Signal transduction

Aspirin, ubiquitin and cancer

Keith D. Wilkinson

Tumours sidestep the body's rules and regulations in myriad ways. The latest discovery ties together a benign tumour syndrome with two hot subjects of biological research: the proteins NF- κ B and ubiquitin.

There are few fields of biology as complicated as the study of cancer. The alphabet soup of abbreviations and acronyms challenges our understanding of an already bewildering array of aberrant cellular signalling pathways. The work of Brummelkamp *et al.*¹, Kovalenko *et al.*² and Trompouki *et al.*³, reported in this issue, adds to the alphabet. The authors find that the tumour-suppressor protein CYLD — loss of which causes a benign human syndrome called cylindromatosis — is a negative regulator of the well-known NF- κ B pathway. Importantly, anti-inflammatory drugs as simple as aspirin could perhaps be an effective intervention.

People with cylindromatosis (also known as turban tumour syndrome) are predisposed to benign tumours that arise in hair follicles and in cells of the sweat and scent glands. The syndrome is caused by a variety of mutations in the CYLD gene, found on chromosome 16 (refs 4, 5). As with all classical tumour suppressors, both copies of the gene must be inactivated to produce cylindromatosis, usually by a mutation in

one copy and a deletion of the region of the chromosome carrying the other copy.

But how does the loss of CYLD cause this tumour syndrome? Many cancers exhibit disturbances in signalling pathways that activate the NF- κ B protein, probably related to this protein's role in regulating genes that control inflammation, the immune response, or cell death⁶. These pathways are themselves regulated at several steps by polyubiquitination — the covalent attachment of a target protein to chains of the small protein ubiquitin⁷. Such chains can have different effects, depending on how they are formed. When the ubiquitins are linked to each other through the lysine amino acid found at position 48 of each ubiquitin, the target protein is directed to the cellular waste-disposal unit, the proteasome⁸. If lysine 63 is used instead, it can serve as a signal for the target to assemble with other proteins^{9,10}.

For various reasons^{5,11}, CYLD has been suspected of being a deubiquitinating enzyme — an enzyme that removes ubiquitin from target proteins. Might it therefore

contribute to the NF- κ B pathway? The new papers^{1–3} show that it does. Brummelkamp *et al.*¹ used a panel of small interfering RNAs to reduce the expression of CYLD and other deubiquitinating enzymes in cultured cells. They found that blocking CYLD expression resulted in the activation of signalling from tumour-necrosis factor- α (TNF- α) through NF- κ B. They also demonstrated a direct interaction between CYLD and NEMO, a component of the NF- κ B pathway. In independent studies, Kovalenko *et al.*² and Trompouki *et al.*³ also identified CYLD as a protein that interacts with NEMO.

So how does NEMO fit into the picture? NF- κ B is normally found in the cell cytoplasm, where it is kept inactive by association with inhibitor (I κ B) proteins. Upon stimulation by proteins such as TNF- α or interleukins, a family of TNF receptors signals to intracellular adapter proteins known as TRAFs (Fig. 1). These proteins have a variety of effects, one of which is to activate the I κ B kinase (IKK) complex. IKK in turn adds phosphate groups to (phosphorylates) I κ B, leading to its polyubiquitination and degradation. NF- κ B is then released from its captor and is free to move into the nucleus, where it regulates a variety of genes. NEMO is one component of the IKK complex, and Kovalenko *et al.*² and Trompouki *et al.*³ show that it binds to the middle of CYLD. Kovalenko *et al.* also found that another part of CYLD can bind to TRAF2. So, it seems likely that TRAF2, NEMO and CYLD are all components of a signalling complex formed in response to TNF- α (although it is not clear whether they can all be part of the complex at the same time).

What does CYLD do there? The new papers^{1–3} show that signalling through the NF- κ B pathway increases when CYLD activity is lost, whether as a result of reducing the protein's concentration with small interfering RNAs, transfecting cells with forms of the protein that bear mutations seen in cylindromatosis patients, or transfecting cells with engineered inactive CYLD. Overexpressing the protein has the converse effect. The findings imply that this protein usually down-regulates signalling, at a step upstream of IKK activation — possibly at the point where TRAFs and NEMO interact.

So what are CYLD's substrates? The papers show that CYLD can deubiquitinate TRAF2 *in vivo*^{1–3} and *in vitro*^{2,3}. It is known that TRAF2 can be connected to ubiquitin chains linked through either lysine 48 or lysine 63; the former leads to its degradation¹² and the latter (which occurs by self-ubiquitination) seems to lead to the assembly of a multi-protein complex that activates IKK, and hence NF- κ B. Kovalenko *et al.* now show that the substrates for CYLD *in vivo* are not lysine-48-linked chains, while Trompouki *et al.* find that the formation of most of the ubiquitinated TRAF2 in the cells

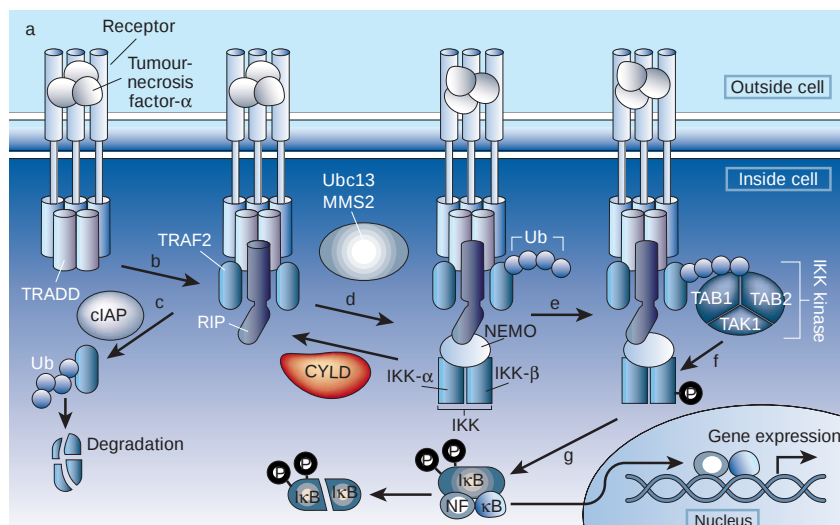


Figure 1 The well-characterized NF- κ B signalling pathway, and a new contributor. **a**, When ligand (such as tumour-necrosis factor- α , TNF- α) binds to TNF receptors, the receptors form a trimer and recruit the TRADD protein. **b**, Subsequently, RIP and TRAFs are recruited. **c**, The cIAP protein can add a chain of ubiquitin proteins, linked through the amino-acid residue lysine 48, to TRAF2, resulting in its degradation. **d**, Alternatively, TRAF2 can ubiquitinate itself, using the Ubc13–MMS2 complex, producing a lysine-63-linked ubiquitin chain. (The availability of cIAP, Ubc13 and MMS2 might determine which pathway predominates.) **e–g**, The IKK kinase complex is recruited (**e**) and phosphorylates IKK (**f**); this in turn phosphorylates I κ B (**g**), leading to its degradation. NF- κ B is then free to move into the nucleus and activate target genes. Three new papers^{1–3} show that the CYLD protein, which is mutated in human cylindromatosis, probably removes lysine-63-linked ubiquitins from TRAF2, thereby preventing further signalling.

studied requires an enzyme of the lysine-63-ubiquitination pathway.

The likely model, then, is that — by removing lysine-63-linked ubiquitin chains from TRAF2 — CYLD causes the dissociation of a complex that would otherwise activate NF- κ B signalling. In the absence of CYLD, as in people with cylindromatosis, NF- κ B signalling is more persistent than normal, contributing to aberrant cell growth. Many other related proteins have deubiquitinating activity^{11,13}, including a protein that is already known to inhibit NF- κ B signalling¹⁴. It is plausible that all these proteins perform the same task in different tissues.

But how could persistent activation of the NF- κ B pathway, which is best known for its role in inflammation, lead to tumours? One hint comes from the observation that activation of the pathway prevents cell death by apoptosis¹⁵. Apoptosis kills cells that have accumulated mutations or other damage — damage that might otherwise lead to cancer. Accordingly, Brummelkamp *et al.* found that a loss of CYLD decreased apoptosis, suggesting that cells lacking this protein could be more susceptible to the accumulation of potentially cancer-predisposing genetic alterations.

Of course questions remain, about both the NF- κ B pathway and the new findings. How, for instance, does the lysine-63-linked ubiquitination of TRAFs lead to the activation of IKK? What is the interplay between the direct and indirect effects of NF- κ B on

cell growth and apoptosis? And can we use this knowledge to develop drugs that ameliorate cylindromatosis and other cancers? Velcade, a proteasome inhibitor that blocks NF- κ B activity, has been approved by the US Food and Drug Administration for the treatment of multiple myeloma¹⁵. Moreover, Brummelkamp *et al.*¹ have shown that aspirin and prostaglandin A1 — anti-inflammatory drugs with effects on the NF- κ B pathway — prevent the activation of NF- κ B-responsive genes and the inhibition of apoptosis caused by the loss of CYLD. Much remains to be investigated, but this finding points to the possibility of a simple, yet potentially important, treatment for cylindromatosis. ■

Keith D. Wilkinson is in the Department of Biochemistry, 4017 Rollins Research Building, 1510 Clifton Road, Emory University, Atlanta, Georgia 30322, USA.

e-mail: genekdw@emory.edu

- Brummelkamp, T. R., Nijman, S. M. B., Dirac, A. M. G. & Bernards, R. *Nature* **424**, 797–801 (2003).
- Kovalenko, A. *et al.* *Nature* **424**, 801–805 (2003).
- Trompouki, E. *et al.* *Nature* **424**, 793–796 (2003).
- Biggs, P. J. *et al.* *Nature Genet.* **11**, 441–443 (1995).
- Bignell, G. R. *et al.* *Nature Genet.* **25**, 160–165 (2000).
- Chen, G. & Goeddel, D. V. *Science* **296**, 1634–1635 (2002).
- Ben-Neriah, Y. *Nature Immunol.* **3**, 20–26 (2002).
- Amit, S. & Ben-Neriah, Y. *Semin. Cancer Biol.* **13**, 15–28 (2003).
- Wang, C. *et al.* *Nature* **412**, 346–351 (2001).
- Deng, L. *et al.* *Cell* **103**, 351–361 (2000).
- Borodovsky, A. *et al.* *Chem. Biol.* **9**, 1149–1159 (2002).
- Li, X., Yang, Y. & Ashwell, J. D. *Nature* **416**, 345–347 (2002).
- Evans, P. C. *et al.* *J. Biol. Chem.* **278**, 23180–23186 (2003).
- Zhang, S. Q., Kovalenko, A., Cantarella, G. & Wallach, D. *Immunity* **12**, 301–311 (2000).
- Richardson, P. *Cancer Treat. Rev.* **29** (Suppl. 1), 33–39 (2003).

Immunology

Who turned on the T cells?

Immunity **19**, 47–57 (2003)

Dendritic cells pick up, internalize and grind up any foreign matter (antigens) that they encounter in the body. Samples of the collected items are displayed at the cell surface for scrutiny by T cells, which then become active. Different tissues have distinct populations of dendritic cells, which generally travel to the lymph nodes to activate T cells there.

One might expect that if an antigen is injected into, for instance, the skin, the dendritic cells there would be the first to activate T cells. Surprisingly, however, Andrea A. Itano and colleagues show that T cells are first turned on by the dendritic cells that pick up the 'skin' antigen after it has been carried by lymph vessels into the lymph node.

Itano *et al.* created a model in which antigens, dendritic cells and antigen-specific T cells could all be monitored in mice. They showed that an initial *ménage à trois* in the lymph node results in proliferation of T cells and secretion of inflammatory proteins (cytokines). Only much later do dendritic cells from the skin arrive to present the antigen to the — by now active — T cells. But this encounter appears to be neither too little nor too late: it is essential for sustained cytokine production and for making T cells that can react swiftly to the same antigen in future.

Marie-Thérèse Heemels

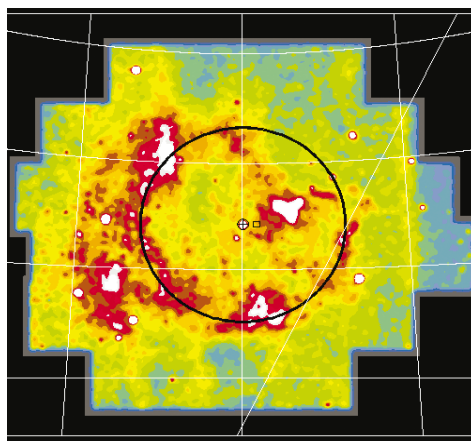
Astrophysics

Pulsar pinpointed

Astrophys. J. **592**, L71–L73 (2003); *Astrophys. J.* **593**, L89–L92 (2003)

By settling a dispute over the distance to a pulsar, Walter F. Briskin and colleagues may have also solved the mystery of a bump in the cosmic-ray spectrum. The observed excess of cosmic rays at an energy of about 3×10^{15} eV might come from the remnant of a supernova explosion that occurred 100,000 years ago, 1,000 light years away. But there was no clear candidate for such a source. This, however, is precisely the place and time estimated by Briskin *et al.* for the supernova that created pulsar PSR B0656+14, in the constellation of Gemini.

This pulsar lies at the centre of an expanding circular shell of supernova debris that reaches between Gemini and the neighbouring constellation Monoceros, called the Monogem ring (pictured). Although the distance to the supernova remnant is measured to be about 1,000 light years, an indirect estimate of the pulsar's distance put it at 2,500 light



The Monogem ring, seen through its X-ray emission. The cross marks the current location of the pulsar PSR B0656+14; its estimated position at birth is indicated by the black rectangle. Both are close to the centre of the expanding ring of gas that is the supernova remnant.

years — suggesting that the superposition is coincidental.

But by making accurate measurements of the parallax of PSR B0656+14 using the Very Long Baseline Array, the researchers have redetermined the pulsar's distance directly. They find it to be 950 light years away.

Philip Ball

Sensory transduction

Not to be sneezed at

Proc. Natl Acad. Sci. USA **100**, 8981–8986 (2003)

Respiratory reflexes such as sneezing help to prevent potentially harmful airborne substances from entering the body. These protective reflexes are triggered by irritation of the trigeminal nerve, and lipid-soluble irritants can diffuse through the lining of the nasal cavity to reach the trigeminal nerve endings. But many compounds that induce respiratory reflexes are insoluble in lipids, so how do these substances tickle the trigeminal nerve?

Thomas E. Finger and colleagues propose an answer: they have identified a new type of receptor cell in the nasal cavity of rats and mice. This cell is similar to solitary chemosensory cells (SCCs), which occur in the skin of non-mammalian aquatic animals. The rodent SCC-like cells carry receptor proteins on their surface that detect bitter substances, such as the fungicide cycloheximide, which is known to irritate the respiratory tract. The chemosensory cells are innervated by the trigeminal nerve, so they probably communicate directly with the nervous system.

This is the first definitive description of an SCC-like chemoreceptor cell in adult mammals, and it will be interesting to see whether these cells are also present in the human nose. If so, it should help

us to understand why so many different stimuli make us sneeze.

Heather Wood

Materials chemistry

Diamonds from dry ice

J. Am. Chem. Soc. **125**, 9302–9303 (2003)

Diamonds up to a quarter of a millimetre across have been made by an entirely new route: chemical reduction of carbon dioxide. Zhengsong Lou *et al.* heated solid CO_2 to 440 °C at a pressure of 800 atmospheres in a sealed container with sodium metal. In the reaction products they found irregular but crystalline diamond particles of about 100 μm or so, along with octahedral crystallites about 10 μm in size.

As well as being produced industrially by high-temperature compression of carbon, diamond can be grown at low pressures by chemical-vapour deposition from a carbon-rich gas, typically a hydrocarbon. This generates polycrystalline films with typical crystal sizes of 0.1–10 μm . The reductive method of Lou *et al.* provides an alternative low-pressure route that is relatively cheap, uses quite mild conditions, and is apparently capable of making surprisingly large crystallites. Graphite (the stable form of carbon under these conditions) is the other major product, but the conversion ratio from CO_2 to diamond can be as high as 8.9 %.

Philip Ball

Environmental biology

Metallic worms turn

Proc. Natl Acad. Sci. USA doi:10.1073/pnas.1731446100

Cadmium-tolerant worms rapidly lose their adaptation when the metal is removed from their environment. The result could lead to new tests for gauging the effectiveness of ecological restoration of contaminated sites.

Jeffrey S. Levinton and colleagues looked at the worm *Limnodrilus hoffmeisteri* living in the seabed of Foundry Cove, New York. Between 1953 and 1979, a battery factory released an estimated 53 tonnes of cadmium into the cove, making it one of the most metal-polluted sites on Earth.

The cove was cleaned up in 1994–95. By 2002, or in about 9–18 worm generations, the researchers found that the creatures were no more resistant to cadmium than those from an unpolluted site. Resistant worms grow slowly, probably because they need to make large quantities of metal-binding protein. In the absence of cadmium, non-resistant worms have an advantage.

This loss of resistance also decreases the potential for cadmium to move up the food-chain. Tracking resistant organisms could therefore be a good test of the health of polluted sites, Levinton and colleagues suggest.

John Whitfield

Bacterial photosynthesis genes in a virus

A bacteriophage may protect itself and its host against a deadly effect of bright sunlight.

Cyanobacteria contribute to the overall photosynthetic production of oxygen in the oceans, but they are susceptible to infection by viruses and also to photo-inhibition when sunlight is too intense. Here we show that the genomic sequence of one such virus, a bacteriophage known as S-PM2, encodes the D1 and D2 proteins that are key components of one of the photosynthetic reaction centres (photosystem II, PSII), which are crucial sites of damage in photo-inhibition. The presence of this virus, and others like it, in the ocean may ensure that photo-inhibition is prevented in infected cells, allowing photosynthesis to continue and therefore provide the energy needed by the virus for its replication.

Primary production in the oligotrophic regions of the oceans is dominated by the cyanobacterial components of the picophytoplankton organisms *Synechococcus* and *Prochlorococcus*¹. The amounts of visible and ultraviolet solar radiation in oceanic ecosystems may be sufficient, particularly in surface waters, to cause photo-inhibition². The primary cause of photo-inhibition in chloroplasts and cyanobacteria is damage to PSII, a large protein–pigment complex that catalyses the light-dependent oxidation of water to molecular oxygen.

At the core of PSII is a dimer of two related proteins, D1 and D2, which binds the pigments and co-factors necessary for the complex's primary photochemistry. During photosynthesis, D1 and, to a lesser extent, D2 turn over rapidly as a result of light-induced damage and are replaced by newly synthesized polypeptides in a repair cycle. When the rate of photo-inactivation and damage of D1 exceeds the capacity for repair, photo-inhibition occurs, resulting in a reduction in the maximum efficiency of PSII photochemistry³.

Viruses in general, and bacteriophages (viruses that infect bacteria) in particular, are abundant in marine ecosystems and are thought to exert major biogeochemical and ecological effects on the marine environment⁴. We analysed the genome sequence of S-PM2 (Fig. 1), a bacteriophage that infects marine *Synechococcus* strains⁵ and is about 194 kilobases long⁶, and found that it encodes the D1 and D2 proteins of PSII.

The genome contains a region of about 3.8 kilobases (GenBank accession no. AY329638) that extends from just over 100 base pairs upstream of the D1 gene (*psbA*), through two genes that are unrelated to photosynthesis, to a point some 100 base pairs downstream of the carboxy terminus of the D2 gene (*psbD*). The two intervening genes encode a homologue of the bacteriophage T4 gp49

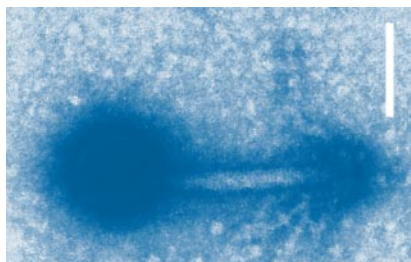


Figure 1 The bacteriophage S-PM2 (here artificially coloured blue), which infects marine cyanobacteria. Scale bar, 100 nm.

(also known as recombination endonuclease VII) and a protein that has some similarity to a putative membrane protein present in the bacterium *Escherichia coli*.

The *psbA* gene appears to be interrupted, as the translated product of the gene aligns with other cyanobacterial and plant D1 proteins up to residue 276 (see supplementary information). There is then a region of 212 base pairs that encodes an amino-acid sequence without any similarity to known D1 proteins; this is followed by a region that encodes the remaining 25 amino acids of D1. This suggests that the *psbA* gene in S-PM2 contains a self-splicing intron. Introns occur in the *psbA* genes of the protozoan *Chlamydomonas*, and there is evidence for light/redox-regulated splicing of *psbA* precursor-messenger RNAs⁷. We detected copies of *psbA* genes after amplification by the polymerase chain reaction in five out of eight other *Synechococcus* viruses, although these genes seem to lack the putative intron.

The complete D1 protein of S-PM2 is similar to the D1 proteins of the marine *Synechococcus* sp. WH8102 (see supplementary

information). There is homology in the DNA sequences, indicating that S-PM2 might have acquired the gene horizontally from its *Synechococcus* host. Presumably, *psbD* was acquired independently, given the presence of two unrelated intervening genes.

The expression of virus-encoded D1 and D2 proteins in infected cells would allow a repair cycle to operate in PSII after the host's protein synthesis had been shut down, thereby maintaining the cells' photosynthetic activity and the concomitant evolution of oxygen, and ensuring the provision of energy for the continued replication of the virus. This survival strategy resembles one used by a virus that infects the green alga *Chlorella*, which enhances the mechanism used by the host cell to rid itself of surplus light energy to avoid photo-inhibition⁸.

Nicholas H. Mann, Annabel Cook, Andrew Millard, Shaun Bailey, Martha Clokie

Department of Biological Sciences, University of Warwick, Coventry CV4 7AL, UK

e-mail: n.h.mann@warwick.ac.uk

- Partensky, F., Blanchot, J. & Vaulot, D. in *Marine Cyanobacteria* (eds Charpy, L. & Larkum, A. W. D.) Vol. 19, 457–475 (Institut Océanographique, Monaco, 1999).
- Piazana, H., Perez-Rodriguez, E., Häder, D.-P. & Lopez-Figueroa, F. *Deep Sea Res. II* **49**, 3513–3528 (2002).
- Aro, E. M., Virgin, I. & Andersson, B. *Biochim. Biophys. Acta* **1143**, 113–134 (1993).
- Fuhrman, J. A. *Nature* **399**, 541–548 (1999).
- Wilson, W. H., Joint, I. R., Carr, N. G. & Mann, N. H. *Appl. Environ. Microbiol.* **59**, 3736–3743 (1993).
- Hambly, E. et al. *Proc. Natl Acad. Sci. USA* **98**, 11411–11416 (2001).
- Deshpande, N. N., Bao, Y. & Herrin, D. L. *RNA* **3**, 37–48 (1997).
- Seaton, G. G. R., Hurry, V. M. & Rohozinsky, J. *FEBS Lett.* **389**, 319–323 (1996).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

COMMUNICATIONS ARISING

Cell signalling

Cell survival and a Gadd45-factor deficiency

Tumour-necrosis factor- α (TNF- α) is an important protein that regulates inflammation, immunity, cell survival and cell death in response to infection or chronic stress^{1–5}. De Smaele *et al.* report that the gene encoding an inducible cellular factor, Gadd45 β (for 'growth arrest and DNA damage'), is essential for promoting TNF- α -mediated cell survival⁶. However, we show here that neither TNF- α signalling nor cell survival is affected in mice lacking *gadd45 β* , a fact which demonstrates that genes other than *gadd45 β*

(refs 5, 7) might regulate cell survival in response to TNF- α .

TNF- α concomitantly activates two signalling pathways — a cell-survival pathway regulated by the transcription factor NF- κ B and a cell-death pathway associated with the c-Jun N-terminal kinase (JNK) cascade^{1–3,5}. Understanding the molecular basis for the cell's choice between life and death in response to TNF- α is a topic of intense investigation^{4,5}. The fate of a cell exposed to TNF- α is determined by cross-talk between these two signalling pathways^{4–7}. NF- κ B activation switches on genes that blunt the pro-cell-death activity of the JNK pathway^{4,6,7}.

De Smaele *et al.* propose that one such NF- κ B target that is essential in TNF- α -mediated cell survival is the *gadd45 β* gene⁶. *Gadd45 β* (also termed *MyD118*) encodes

one of the family of Gadd45 proteins, which are implicated in growth arrest, DNA-damage repair and programmed cell death^{8,9}. De Smaele *et al.* showed that ectopic over-expression of Gadd45β in mouse-embryo fibroblasts (MEFs) and in NF-κB-deficient cell lines antagonizes TNF-α-induced cell death and increases cell survival. In addition, ectopic expression of antisense *gadd45β* messenger RNA, which presumably blocks Gadd45β expression, was found to decrease cell survival and prolong JNK activity; this is similar to the response in cells lacking NF-κB. The authors conclude⁶ that TNF-α-mediated activation of NF-κB induces Gadd45β, which inhibits TNF-α-mediated cell death and JNK signalling and promotes cell survival; however, this conclusion contradicts an earlier study¹⁰ that implicates Gadd45β as an activator of JNK.

We used *gadd45β*-null mice, in which the *gadd45β* gene is ablated (D.L. and A.F., unpublished results), to assess further the effect of *gadd45β* deficiency on TNF-α-mediated cellular responses, including cell survival and JNK signalling. We found that TNF-α induced *gadd45β* expression in wild-type but not in *gadd45β*-deficient MEFs (Fig. 1a). Like wild-type MEFs, *gadd45β*-deficient (*gadd45β*^{-/-}) MEFs were not susceptible to TNF-α-mediated cell death (Fig. 1b). However, in the presence of the protein-synthesis inhibitor cycloheximide, which prevents the expression of the pro-survival genes induced by NF-κB, both *gadd45β*^{-/-} and wild-type MEFs were equally susceptible to TNF-α-mediated cell death.

Our findings indicate that *gadd45β* expression is not essential for the NF-κB pro-survival function. Furthermore, the kinetics

of downregulation of JNK activity were similarly rapid in *gadd45β*^{-/-} MEFs and in wild-type cells (Fig. 1c), as well as in another cell type, splenic lymphocytes (data not shown).

Our results indicate that other NF-κB target genes^{5,7} are more likely than *gadd45β* to be primary mediators of the survival function of NF-κB. The discrepancy between our observations and those of De Smaele *et al.*⁶ might reflect limitations in their experimental approach — for example, ectopic over-expression of *gadd45β* or of its antisense RNA in cells stimulated with TNF-α might have affected cell survival and JNK activity in some indirect or nonspecific way. Further work is needed to assess what role, if any, Gadd45β has in the cell's response to TNF-α.

Arshad Amanullah*†, Naiyer Azam*, Arthur Balliet*, Christine Hollander‡, Barbara Hoffman*, Albert Fornace Jr‡, Dan Liebermann*

*Fels Institute for Cancer Research and Molecular Biology and Department of Biochemistry, Temple University School of Medicine, Philadelphia, Pennsylvania 19140, USA
e-mail: lieberma@unix.temple.edu

‡Gene Response Section, Center for Cancer Research, National Cancer Institute, NIH, Bethesda, Maryland 20892-4255, USA

†Present address: HTB/Integrative Biology Initiative, GlaxoSmithKline, King of Prussia, Pennsylvania 19406, USA

1. Baud, V. & Karin, M. *Trends Cell Biol.* **11**, 372–377 (2001).
2. Locksley, R. M., Killeen, N. & Lenardo, M. J. *Cell* **104**, 487–501 (2001).
3. Chen, G. & Goeddel, D. V. *Science* **296**, 1634–1635 (2002).
4. Kyriakis, J. M. *Nature* **414**, 265–266 (2001).
5. Karin, M. & Lin, A. *Nature Immunol.* **3**, 221–227 (2002).
6. De Smaele, E. *et al.* *Nature* **414**, 308–313 (2001).
7. Tang, G. *et al.* *Nature* **414**, 313–317 (2001).
8. Liebermann, D. A. & Hoffman, B. *Leukemia* **16**, 527–541 (2002).
9. Fornace, A. J. Jr, Jackman, J., Hollander, M. C.,

Hoffman-Liebermann, B. & Liebermann, D. A. *Ann. NY Acad. Sci.* **663**, 139–153 (1992).

10. Takekawa, M. & Saito, H. *Cell* **95**, 521–530 (1998).

De Smaele *et al.* reply — We and others have shown that the control of TNF-α-induced apoptosis by NF-κB/Rel transcription factors involves suppression of the JNK enzyme cascade^{1–3}, and we have proposed that this suppression is mediated in part by Gadd45β/Myd118 (refs 1,4). Amanullah *et al.* suggest that the ablation of *gadd45β* has no effect on JNK activation and apoptosis by TNF-α and argue that the protective effects of NF-κB are mediated by factors other than Gadd45β.

However, caution is needed in drawing inferences from these provocative findings about the role of Gadd45β in the cell. Under the conditions used by Amanullah *et al.*, knockout mutation of any of the NF-κB targets identified so far^{5,6} — including those of the putative JNK inhibitor XIAP (ref. 7) and of NF-κB/RelA itself⁸ (our unpublished observations) — would not have affected TNF-α-induced killing. This is because cytokine treatment of fibroblasts was far too short and was performed in the absence of low doses of cycloheximide (about 0.1 μg ml⁻¹; ref. 1), which is needed to down-regulate functionally redundant factors.

Our antisense experiments¹ indicate that in certain cells, such as lymphoid cell lines, downregulation of *gadd45β* leads to exaggerated JNK signalling and apoptosis in response to TNF-α. It is likely that the pro-survival programme that is activated by NF-κB depends on tissue-specific elements^{5,6}, so the relevance of Gadd45β to this protective activity of NF-κB might be more marked in certain cell types. As the analysis of Amanullah *et al.* is limited to the fibroblastoid lineage, it might not be appropriate to generalize conclusions about the effects of Gadd45β on the JNK pathway and apoptosis to other cell types. We agree with Amanullah *et al.* that further investigation is needed to define the precise contribution of this factor and of other targets to the anti-apoptotic function of NF-κB.

Francesca Zazzeroni, Salvatore Papa, Enrico De Smaele*, Guido Franzoso

Ben May Institute for Cancer Research, and Gwen Knapp Center for Lupus and Immunology Research, University of Chicago, Chicago, Illinois 60637, USA
e-mail: gfranzos@midway.uchicago.edu

*Present address: Department of Experimental Medicine and Pathology, University of Rome 'La Sapienza', 00161 Rome, Italy

1. De Smaele, E. *et al.* *Nature* **414**, 308–313 (2001).
2. Tang, G. *et al.* *Nature* **414**, 313–317 (2001).
3. Javelaud, D. & Besancon, F. *Oncogene* **20**, 4365–4372 (2001).
4. Franzoso, G., Zazzeroni, F. & Papa, S. *Cell Death Diff.* **10**, 13–15 (2003).
5. Baldwin, A. S. J. *Clin. Invest.* **107**, 241–246 (2001).
6. Karin, M. & Lin, A. *Nature Immunol.* **3**, 221–227 (2002).
7. Harlin, H., Reffey, S. B., Duckett, C. S., Lindsten, T. & Thompson, C. B. *Mol. Cell. Biol.* **21**, 3604–3608 (2001).
8. Beg, A. A. & Baltimore, D. *Science* **274**, 782–784 (1996).

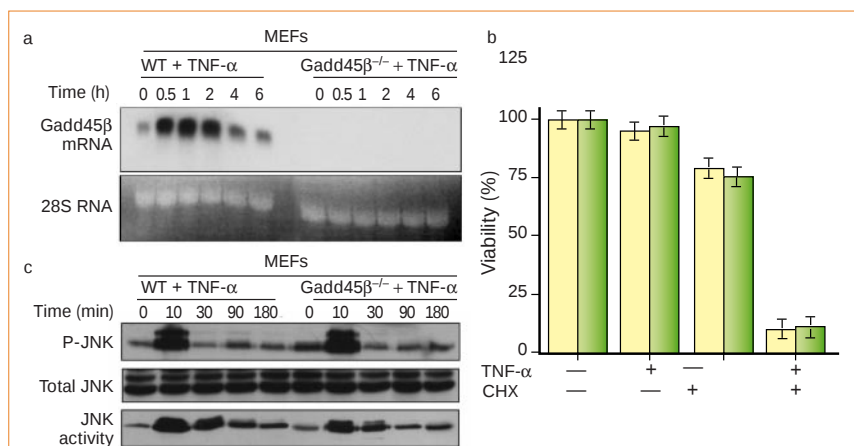


Figure 1 Gadd45β deficiency does not alter TNF-α-mediated cell survival or JNK signalling. **a**, TNF-α induces expression of *gadd45β* in wild-type (WT) but not in *gadd45β*^{-/-} mouse-embryo fibroblasts (MEFs). RNA from cells collected at different times after treatment with TNF-α (1,000 U ml⁻¹) was resolved on a 2% agarose-formaldehyde gel and analysed by northern blotting using a ³²P-labelled *gadd45β* probe. 28S RNA, loading control. **b**, *gadd45β*^{-/-} cells are not susceptible to TNF-α-mediated cell death in the absence of cycloheximide (CHX). Viability of WT (left bars) and *gadd45β*^{-/-} MEFs (right bars) was determined by trypan-blue staining after 14-h treatment with TNF-α (1,000 U ml⁻¹), CHX (1 mg ml⁻¹) or both. **c**, Gadd45β is not required for rapid downregulation of JNK after TNF-α treatment. MEFs from WT and *gadd45β*^{-/-} mice were treated with TNF-α (1,000 U ml⁻¹) and, at the indicated times, cell lysates were resolved by SDS-PAGE and western blotted (antibody probes: antiphospho-JNK, anti-JNK; Cell Signaling, Inc). A fusion protein of glutathione-S-transferase and c-Jun was used as substrate in the JNK assay (bottom).

Identification of *Lps2* as a key transducer of MyD88-independent TLR signalling

K. Hoebe¹, X. Du¹, P. Georgel¹, E. Janssen², K. Tabeta¹, S. O. Kim¹, J. Goode¹, P. Lin¹, N. Mann¹, S. Mudd¹, K. Crozat¹, S. Sovath¹, J. Han¹ & B. Beutler¹

¹Department of Immunology, IMM-31, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA

²Division of Cellular Immunology, La Jolla Institute for Allergy and Immunology, 10355 Science Center Drive, San Diego, California 92121, USA

In humans, ten Toll-like receptor (TLR) paralogues sense molecular components of microbes, initiating the production of cytokine mediators that create the inflammatory response. Using *N*-ethyl-*N*-nitrosourea, we induced a germline mutation called *Lps2*, which abolishes cytokine responses to double-stranded RNA and severely impairs responses to the endotoxin lipopolysaccharide (LPS), indicating that TLR3 and TLR4 might share a specific, proximal transducer. Here we identify the *Lps2* mutation: a distal frameshift error in a Toll/interleukin-1 receptor/resistance (TIR) adaptor protein known as Trif or Ticam-1. *Trif*^{*Lps2*} homozygotes are markedly resistant to the toxic effects of LPS, and are hypersusceptible to mouse cytomegalovirus, failing to produce type I interferons when infected. Compound homozygosity for mutations at *Trif* and *MyD88* (a cytoplasmic TIR-domain-containing adaptor protein) loci ablates all responses to LPS, indicating that only two signalling pathways emanate from the LPS receptor. However, a *Trif*-independent cell population is detectable when *Trif*^{*Lps2*} mutant macrophages are stimulated with LPS. This reveals that an alternative MyD88-dependent 'adaptor X' pathway is present in some, but not all, macrophages, and implies afferent immune specialization.

During the past several years the TLRs have been shown to act as primary innate immune sensors, each responding to specific molecules of microbial origin^{1–7}. They alert the host to the presence of infectious organisms within minutes after inoculation, and also initiate the systemic inflammatory state that accompanies sepsis⁸, marked by the production of tumour-necrosis factor (TNF), interleukin (IL)-1, IL-6, IL-12, type I interferons and other proteins. All TLRs are single-spanning membrane proteins that display a conserved cytoplasmic motif (the TIR domain⁹), and it is thought that MyD88 acts as a transducer of signalling by all TLRs with the possible exception of TLR3. MyD88 signals by way of IRAK-4, a serine kinase required for the effective production of inflammatory cytokines induced by all of the TLRs^{10,11}. But MyD88 is not the sole adaptor for TLR signalling. At least five TIR domain adaptors are encoded in the human genome, and some TLRs depend on more than one adaptor. The endotoxin (lipopolysaccharide; LPS) receptor, TLR4, provides an example.

As mice lacking MyD88 still exhibit some LPS responses, the existence of a 'MyD88-independent' pathway has been proposed¹². One of the hallmarks of this pathway is the activation of interferon- β production¹³, a signalling endpoint also known to be elicited by double-stranded (ds)RNA stimulation¹⁴. Probably as a result of interferon- β receptor stimulation and JAK/STAT activation, the production of nitric oxide (NO)¹² and the chemokine IP-10¹³ are elicited in turn, and may also be considered as elements of the MyD88-independent response. In addition, MyD88-deficient mice retain LPS-activated phosphorylation of mitogen-activated protein kinase (MAPK) family members, including ERK1/2, p38 kinase and Jun kinase (JNK), and show LPS-induced activation of NF- κ B, although each of these events is slightly delayed¹². Although it was suggested that a second cytoplasmic TIR-domain protein, known as Mal (also called Tirap), might serve the MyD88-independent pathway^{15,16}, genetic evidence revealed that Mal and MyD88 act strictly in conjunction with one another^{17,18}. The receptor-proximal components of the MyD88-independent signalling pathway have there-

fore remained uncertain. MyD88-independent signalling might depend on several proteins operating in parallel, or alternatively, might reflect the activity of a single protein or protein complex, carrying signals from TLR4. We have now resolved the molecular basis of MyD88-independent signalling through a forward genetic approach, and herein, we report that this pathway plays a key part in the innate immune response to both viral and bacterial stimuli.

Identification and characterization of the *Lps2* phenotype

In an effort to identify new and essential components of all TLR signalling pathways, we initiated a programme of random germline mutagenesis in mice, using the alkylating agent *N*-ethyl-*N*-nitrosourea (ENU). Both F₁ and F₃ germline mutants were produced on the C57BL/6 background, and males were examined by collecting thioglycolate-elicited peritoneal macrophages under anaesthesia, and testing the integrity of TLR responses *in vitro*. TNF production was measured as an endpoint of signalling in this screen, using the L929 bioassay. LPS, lipid A, peptidoglycan, PAM₃CysSer(Lys)₄, resiquimod, dsRNA and CpG-oligonucleotides were used as inducing agents. A total of 2,649 F₁ mice (carrying heterozygous mutations) and 4,584 F₃ mice (carrying homozygous mutations) have so far been screened for new innate immune phenotypes.

Among these animals, a mouse with strong resistance to lipid A and LPS was identified in the F₃ population¹⁹, and was inbred to generate a homozygous mutant stock. In our initial phenotypic characterization of the mutation, dubbed *Lps2*, we found that not only LPS and synthetic lipid A (which signal through TLR4)¹ but also dsRNA (which signals through TLR3)⁵ were incapable of activating TNF production in macrophages from homozygous mutants (Fig. 1a–c). LPS-induced macrophage cytotoxicity was also abolished by the mutation (Fig. 1h). The mutation had a far weaker effect on IL-12 and IL-6 production (not shown), and the integrity of signalling through TLR1, -2, -7 and -9 was entirely normal (Fig. 1d–g). Zymosan, which signals through TLR2 and TLR6 heteromers, also produced a normal response (not shown).

The MyD88-dependent pathway, as it proceeds in the course of TLR4 activation, involves recruitment of both MyD88 (ref. 12) and Mal^{17,18}, and leads to the phosphorylation of ERK1/2, p38 MAP kinase and JNK, and the transient degradation of I κ B. These events occurred in LPS-activated macrophages from normal and *Lps2* homozygous mutants, although slightly early ERK1/2 phosphorylation, early dephosphorylation and early regeneration of I κ B were consistently evident (Fig. 2a). On the other hand, dsRNA did not stimulate a response in cells from *Lps2* homozygotes (Fig. 2b). The *Lps2* mutation inhibited LPS-induced interferon- β messenger RNA accumulation (Fig. 2c), and type I interferon activity was undetectable in the culture medium of LPS- or dsRNA-activated macrophages from *Lps2* homozygotes (Fig. 2d). LPS-induced NO production as well as inducible NO synthase (iNOS) synthesis, known to be stimulated by type I interferons, were abolished by the mutation (Fig. 2e). LPS- and dsRNA-induced STAT1 phosphorylation was impaired in *Lps2* mutant cells as well, although STAT1 phosphorylation occurred normally in homozygous mutant cells after addition of interferon- β to the culture (Fig. 2f).

All of these effects pointed to proximal blockade of the MyD88-independent pathways serving TLR3 and TLR4. Suspecting a defect

upstream of interferon- β , we examined the formation of phosphorylated IRF-3 dimers in LPS- and dsRNA-activated macrophages by western blotting. Heterozygotes for the *Lps2* mutation showed impaired LPS-induced IRF-3 phosphodimer formation (not shown); homozygotes showed no IRF-3 dimerization (Fig. 2g).

Effects of *Lps2* homozygosity

Despite the apparent integrity of the MyD88–Mal signalling pathway, known to be required for full expression of endotoxicity^{12,17}, *Lps2* homozygotes are markedly resistant to the lethal effect of challenge with a 1-mg dose of *Escherichia coli* LPS (Fig. 3a). It should be noted that all of the mutant animals became ill, and some of them died after LPS challenge. Hence, they were visibly less resistant than mice lacking TLR4^{1,20}. Nonetheless, it is clear that the MyD88-independent pathway contributes to endotoxicity: a finding consistent with the recent demonstration that interferon- β and its signalling apparatus has an important role in LPS-induced shock²¹.

In addition to their faulty response to *E. coli* LPS, *Lps2* homozygotes are compromised in their response to infection with mouse cytomegalovirus (mCMV). Three of eight mutant mice died after challenge with an inoculum (5×10^5 plaque-forming units (PFU)), whereas 30 out of 30 of normal C57BL/6 mice survived, and 1,000-fold higher viral titres were present in the spleens of *Lps2* homozygotes as compared with C57BL/6 controls. BALB/c mice, which fail to eliminate the virus as a result of a mutation in the NK-activating receptor LY49H^{22,23}, were used as positive controls and developed still higher titres of the virus (Fig. 3b). It remains to be determined whether TLR3 or a signalling protein yet to be identified is of critical importance in these responses, but it is clear that *Lps2* is an essential element of a response pathway required for containment of infection.

As the *Lps2* mutation abolished LPS- and dsRNA-induced type I interferon production by macrophages *in vitro*, we examined type I interferon levels in the serum of mCMV-infected mice and in culture medium collected from mCMV-infected macrophages. Robust production of type I interferon was observed in the serum of normal mice, but no type I interferon was detected in the serum of *Lps2* mutant homozygotes (Fig. 3c). The response of macrophage cultures to infection by *Vaccinia* virus was demonstrably impaired, in that macrophage monolayers from *Lps2* homozygotes supported the replication of *Vaccinia* to a higher titre than did macrophages from normal mice (Fig. 3d). Hence, the antiviral effect conferred by the normal *Lps2* product is very broad with regard to the virus species, correlated with type I interferon production, and is at least largely independent of an adaptive immune response. Moreover, the normal *Lps2* product is required for a type I interferon response to viral infection *in vivo*.

Genetic mapping and sequencing reveals the identity of *Lps2*

Lps2 was mapped meiotically by out-crossing homozygotes to C3H/HeN mice and backcrossing the progeny to the mutant stock. Using a panel of 59 genome-wide microsatellite markers applied to 26 meioses, *Lps2* was first confined to an interval on mouse chromosome 17 (Supplementary IS-1), and then, on a total of 1,567 meioses, to a region delimited by microsatellite markers D17Mit87 and D17Mit20. Within this region, eleven additional microsatellite length polymorphisms were identified in our laboratory (Supplementary IS-4). A total of seven crossovers were observed between D17Mit87 and D17Mit20, and these crossovers were resolved by haplotype analysis with respect to three of the internal markers, permitting optimal confinement of the *Lps2* mutation (Supplementary IS-2). A single crossover was observed between the proximal marker *Lps2*_7_28 and *Lps2*. A single crossover was also observed between *Lps2* and the distal marker *Lps2*_7_34. The length of the critical region, thus defined, was 216 kilobases (kb). The critical region was amplified in 63 overlapping 4-kb fragments using template amplified from an *Lps2*

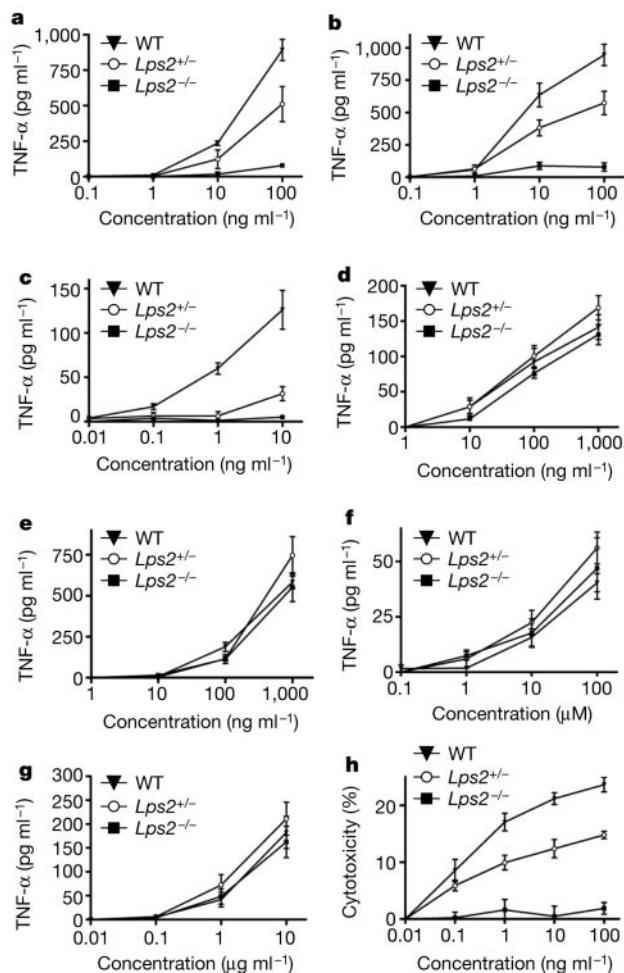


Figure 1 TLR3 and TLR4 specificity of the *Lps2* mutation. Mice were injected intraperitoneally with 3% thioglycolate. After 4 days macrophages were isolated and cultured at a density of 5×10^5 cells per well in 96-well plates. Dose-response experiments were performed for each specific inducer. After 4 h of incubation at 37 °C, supernatants were collected and assayed in duplicate for TNF- α activity as described previously³². LPS cytotoxicity was evaluated after 24 h of incubation. Values represent means $\pm$ s.e.m. ($n = 6$ mice). **a**, Lipid A; **b**, LPS; **c**, poly(I:C); **d**, PAM₃CysSer(Lys)₄; **e**, resiquimod; **f**, CpG-containing DNA; **g**, peptidoglycan; **h**, LPS-induced cytotoxicity.

homozygous mutant and from a normal C57BL/6 mouse, and sequenced in entirety on both strands with a panel of 865 internal primers. Only a single mutation was present in the critical region of the *Lps2* strain, and it was identified within one of eight putative genes that reside within the region (Supplementary IS-3a). This candidate, although fragmentary and unannotated in the Ensembl database, was considered exceptionally promising because it encodes a TIR-domain-containing protein. Two reports have previously suggested that this protein, elsewhere called Trif²⁴ or Ticam-1 (ref. 25), might indeed have a role in TLR signalling, and

particularly in the activation of interferon- β gene expression. However, no consensus was established as to the TLRs that depend on it; one report suggesting association with numerous TLRs and MyD88 (ref. 24), the other an exclusive association with TLR3 (ref. 25).

The length and content of the murine *Trif* gene was deduced by analysis of the numerous expressed sequence tags that have been derived from it, and confirmed by rapid amplification of cloned ends (RACE) of macrophage-derived complementary DNA. In normal C57BL/6 mice, the gene encodes a protein 732 residues in

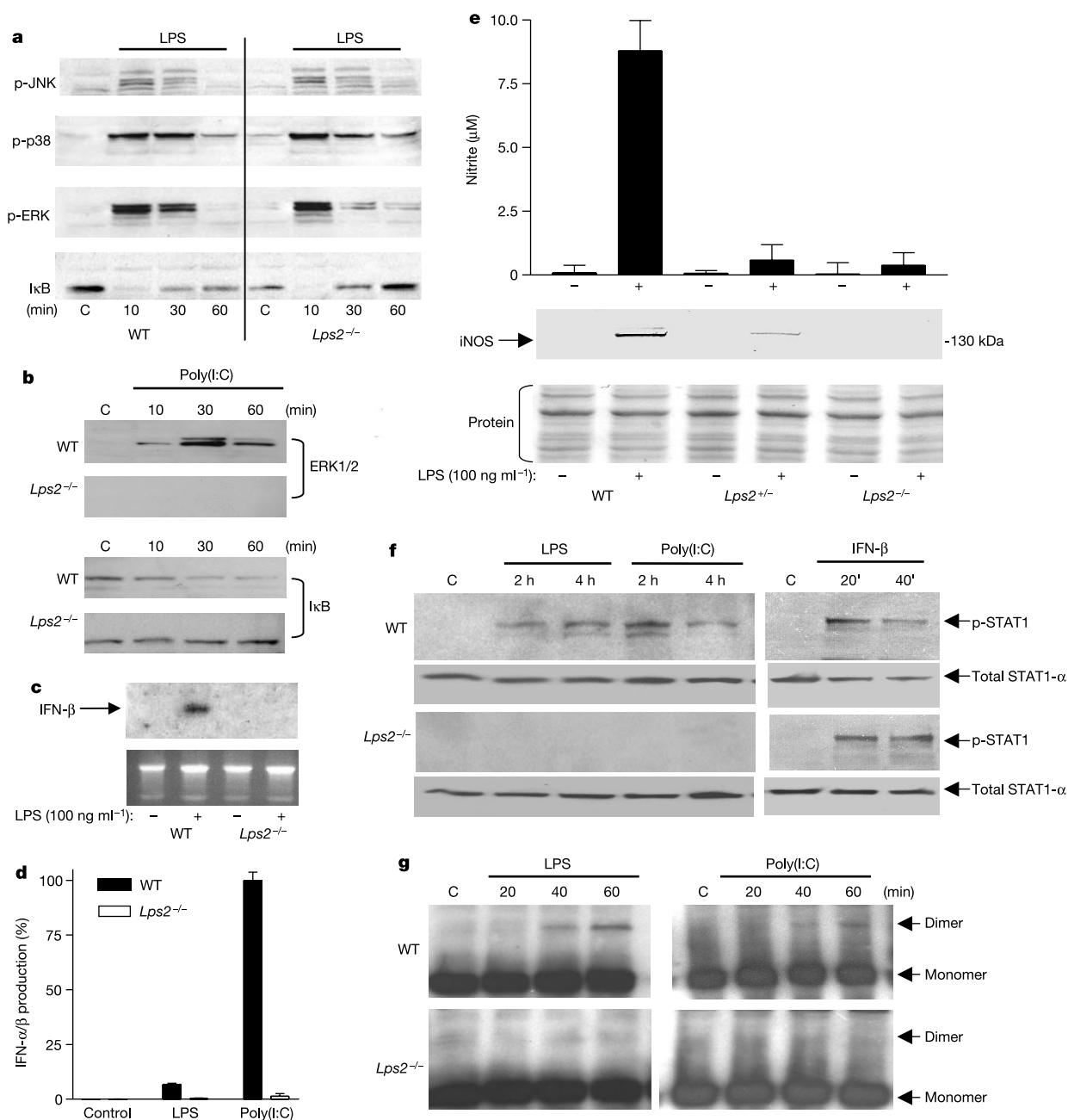


Figure 2 *Lps2* is required for MyD88-independent TLR3 and TLR4 signalling in macrophages. **a**, **b**, Western blot analysis of phospho-MAPKs and I κ B after macrophage stimulation with LPS (**a**) or poly(I:C) (**b**). **c**, Northern blot analysis after 1 h incubation with or without LPS. **d**, Type I interferon activity measured after 2 h incubation with LPS (100 ng ml⁻¹) or poly(I:C) (10 μ g ml⁻¹). Error bars indicate s.e.m.; $n = 2$ mice. **e**, iNOS

(western blot) and NO (Griess assay) were measured after 24 h incubation with or without LPS. Protein was stained with Ponceau S. Error bars indicate s.e.m.; $n = 6$. **f**, **g**, Western blot analysis of STAT1 (**f**) and IRF-3 phosphodimer formation (**g**); LPS, 100 ng ml⁻¹; poly(I:C), 10 μ g ml⁻¹.

length. In *Trif*^{Lps2} homozygotes, the Trif protein is modified by a single base-pair deletion, predicted to remove 24 amino acids from the carboxy terminus of the protein, replacing them with an unrelated sequence 11 amino acids in length (Supplementary IS-3b, c). Because this was the sole mutation present in the *Lps2* critical region, as the mutation markedly alters the *Trif* coding sequence, and as the mutant phenotype is very much consistent with modification of a TIR-domain-containing protein, we concluded that the *Lps2* phenotype results from the observed *Trif* mutation. This conclusion was supported by phenotypic rescue experiments in which macrophages were transfected with an expression vector encoding normal Trif. Transfected cells, identified by gating on fluorescence created by co-transfection with a yellow fluorescent protein (YFP) expression vector, were capable of producing TNF in response to dsRNA (Fig. 4). Moreover, the findings suggest that the mutant product may exert a dominant inhibitory effect.

Notably, the 19 C-terminal residues of the mouse Trif protein are not represented in the human homologue of the protein, which is nonetheless functional in our own transfection system and in those reported elsewhere^{24,25}. Moreover, a large C-terminal deletion of the human protein reportedly failed to impede Trif-mediated signal transduction when expressed in mammalian cells; only a combined amino-terminal/C-terminal deletion ($\Delta N\Delta C$) had dominant inhibitory properties²⁴. These considerations may indicate that the 11-residue aberrant C-terminal sequence created by the *Lps2* frameshift error causes instability or inactivation of the mutant protein. The co-dominant character of the *Trif*^{Lps2} allele would in this context probably be explained on the basis of a quaternary structural effect. The possibility that the locus is also haploinsufficient cannot, as yet, be excluded. A null allele of *Trif* (for example, deletion of the locus) might conceivably yield a distinguishable

phenotype, more consistent with either of the previous *in vitro* studies^{24,25}, but this remains to be established.

A tBLASTn search using the Trif protein as a query retrieves another gene encoding a TIR domain polypeptide, localized to mouse chromosome 18 (46.7 megabases from the centromere; ENSMUST00000036100). This, the closest homologue of Trif, is also represented in the human genome, and seems to be a cytoplasmic adaptor protein similar to Trif itself. As yet, no function has been ascribed to this homologue. We suggest that it may be involved in TIR-related signal transduction, and consider it possible that it also has an essential role in Trif-independent signalling, detailed below.

The bifid nature of the LPS signalling pathway

As described above, TLR3 and TLR4 share a common signalling intermediate, Trif, which is absolutely required for IRF-3 phosphodimer formation occurring in response to LPS, dsRNA or authentic viral infections, and all subsequent events related to type I interferon production and action. Particularly with regard to LPS signalling, however, it should be noted that some of the events that occur in MyD88-deficient mice, such as the tardive phosphorylation of ERK1/2, p38, JNK, and activation of NF- κ B, also occur in homozygous *Trif*^{Lps2} mutants. Hence, two possibilities present themselves. On the one hand, Trif and the MyD88–Mal complex might each be capable of initiating these ‘shared’ endpoints of TLR signalling independently. In the most parsimonious model, all LPS signals might traverse two (and only two) ‘branches’, one relatively rapid and dependent on MyD88 and Mal, the other slightly slower and dependent on Trif. On the other hand, an unknown number of signalling adaptors might be responsible for MAPK family and NF- κ B activation independent of both Trif and the MyD88–Mal complex. In this alternative, the MyD88-independent pathway would not be unitary, but would comprise at least two and perhaps several pathways, each arising independently from the LPS receptor.

To make a clear distinction between these alternatives, we sought to determine whether any signals at all emanate from the activated LPS receptor in *Trif*^{Lps2/Lps2}; *MyD88*^{-/-} mice. To produce compound homozygous mutants, we crossed *MyD88*^{-/-} mice (a gift of

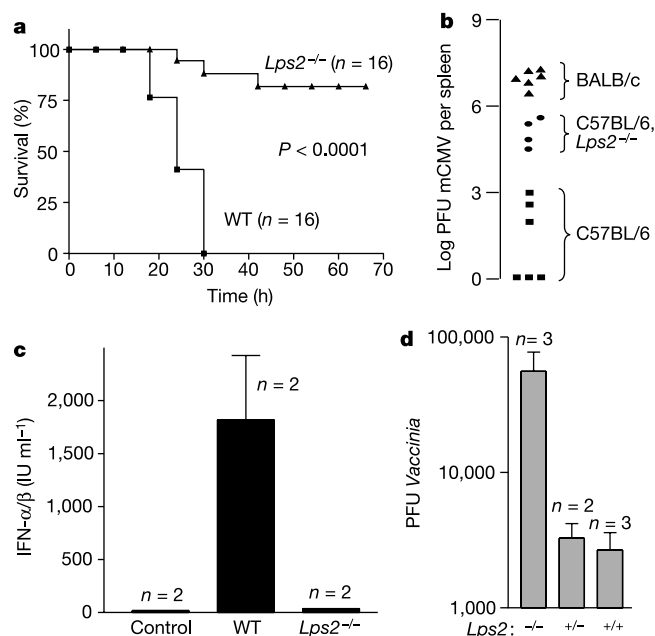


Figure 3 Responses to bacterial and viral stimuli in *Lps2* homozygotes. **a**, Mice were injected intraperitoneally with 1 mg of *E. coli* LPS and monitored for survival. **b**, Mouse cytomegalovirus (mCMV) infection. Each mouse received 5×10^5 PFU intraperitoneally. Spleens were assayed for infectious virus by plaque assay using NIH 3T3 cells at 5 days after inoculation. **c**, Type I interferon activity measured in the serum of mCMV-infected mice³¹, 1.5 days after inoculation. Error bars indicate s.e.m. **d**, 4×10^6 peritoneal macrophages were infected with 4×10^5 PFU of *Vaccinia* virus (titled on HeLa cells); assay of viral product was performed at 3 days after inoculation. *n* indicates the number of mice; error bars indicate 95% confidence limits.

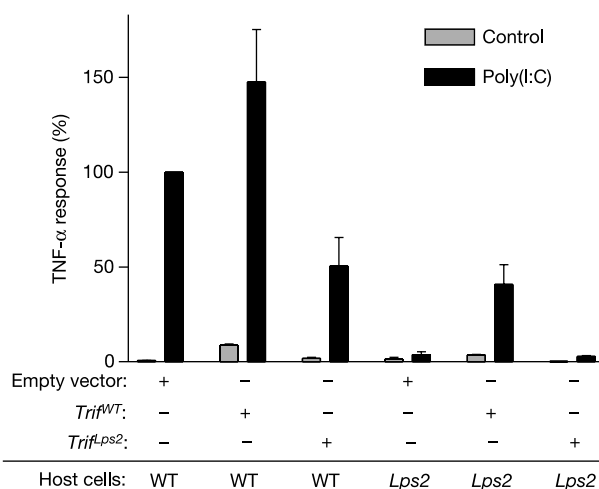


Figure 4 Rescue of poly(I:C) responsiveness in *Lps2* homozygous mutant macrophages by transfection with normal cDNA. Relative quantity of TNF in macrophages from normal and mutant mice, transfected with constructs shown, with or without poly(I:C) induction, $10 \mu\text{g ml}^{-1}$ for 4 h. All values are normalized with respect to the quantity of TNF expressed by induced wild-type cells transfected with empty vector (100%). Error bars indicate s.e.m. of three independent experiments. WT, C57BL/6 macrophages; *Lps2*, macrophages from homozygous mutant mice.

S. Akira) to *Trif*^{Lps2} homozygotes, and intercrossed the offspring. Among 88 F₂ animals examined, three *Trif*^{Lps2/Lps2}; *MyD88*^{-/-} animals were identified when genotyped at three months of age. Peritoneal macrophages obtained from these animals were examined for phosphorylation of MAPK family members and transient degradation of IκB (Fig. 5).

Single-locus mutants both showed normal levels of ERK1/2 phosphorylation and transient degradation of IκB. As previously described¹², *MyD88* deletion causes a slight delay in these events, but does not abrogate them. Homozygosity for the *Trif*^{Lps2} allele resulted in slightly earlier phosphorylation by comparison, with a concomitant early decline in the signal. Compound homozygotes showed no ERK1/2 phosphorylation responses (Fig. 5a), nor transient degradation of IκB (Fig. 5b). Precisely the same pattern of response was observed for p38 kinase and JNK (not shown). Although more extensive studies may ultimately reveal endpoints of LPS signalling that are as yet unknown, it seems for the present that MyD88–Mal signalling and Trif signalling define separate rami of a bifurcated pathway responsible for all of the signals that emanate from the LPS receptor. As it is well known that all LPS signals are silenced by the P732H mutation of TLR4 (ref. 1), which does not seem to alter receptor shape²⁶ and is consistent with the surface expression of TLR4 (ref. 27), it may be inferred that the separate adaptors both engage the same domain of the receptor, and are probably in competition with one another.

Both branches of the LPS signalling pathway are required for the full expression of LPS toxicity *in vivo*, and it is clear that the two branches are superadditive, insofar as integrity of either locus yields

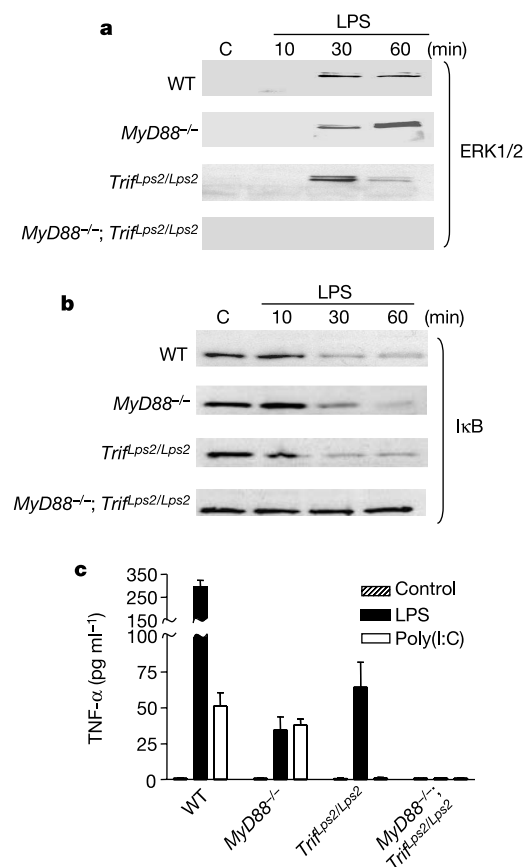


Figure 5 Single and compound homozygosity for *MyD88* and *Trif* mutations: evidence for a bifid LPS response pathway and a unitary dsRNA response pathway. **a–c**, Macrophages were left unstimulated or activated with LPS (10 μg ml⁻¹) and analysed for phosphorylation of ERK1/2 (**a**), IκB (**b**), or TNF-α secretion measured by bioassay (**c**). Each bar represents 2–4 mice; error bars indicate s.e.m.

only about 10% of the amount of TNF produced by wild-type cells (Fig. 5c). Synergy between the MyD88–Mal pathway and the Trif pathway might occur distally as well, through the cooperative elicitation of cytokines and other effector proteins, many of which are transcriptionally activated by both STAT1 and NF-κB. Insofar as LPS-induced macrophage apoptosis is abolished by the *Trif*^{Lps2} mutation (Fig. 1h), we conclude that pro-apoptotic signals emanate exclusively from the Trif pathway. As a double mutation is required to prevent MAPK phosphorylation and IκB degradation, we conclude that both Trif and MyD88–Mal are independently capable of initiating these events, the former more slowly than the latter. Although the consequences of MyD88–Mal signalling are temporally distinguishable from those of the Trif pathway, we have not, as yet, identified a qualitatively unique effect of MyD88–Mal signalling induced by LPS. It is unclear from our own data whether Trif directly engages IRF-3 and/or whether it engages an essential kinase that does so²⁸.

Characteristics of a Trif-independent signalling pathway

Just as it is possible to speak of MyD88-dependent and MyD88-independent signalling, it is also possible to speak of Trif-dependent and Trif-independent signalling, as the mutation that we have observed does not fully abolish LPS signal transduction, although it does fully abolish dsRNA signal transduction. Notably, Trif-dependent and Trif-independent signalling pathways reside within separate cell populations (Fig. 6a, b).

All CD11b-positive, F4/80-positive macrophages in a thioglycolate-elicited peritoneal exudate from wild-type mice respond to LPS (Fig. 6a, top), whereas only a fraction of the cells respond to dsRNA (Fig. 6b, top) by producing TNF. Macrophages from *Trif*^{Lps2/Lps2} mice are nearly unresponsive to dsRNA (Fig. 6b, bottom); however,

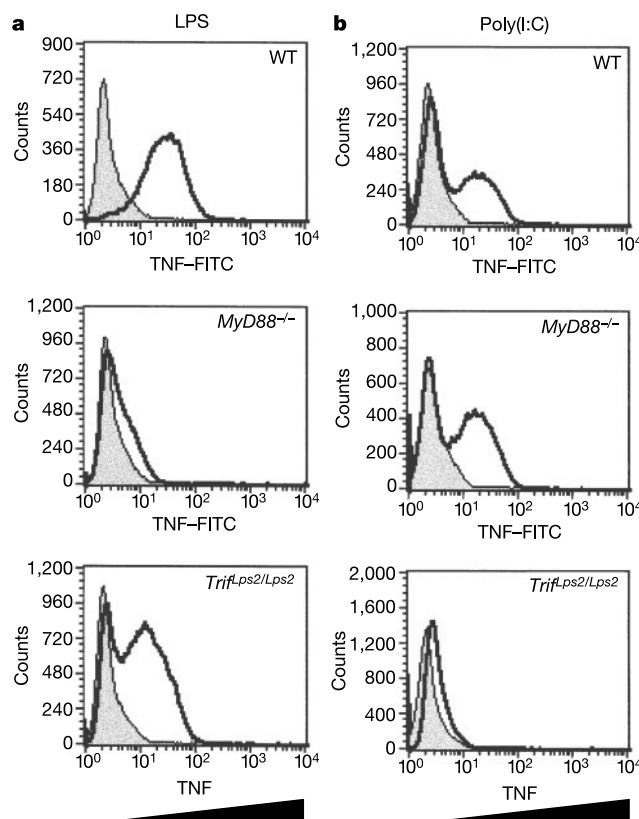


Figure 6 Evidence for functionally distinct populations of macrophages. Distribution of TNF-α accumulation in cells left unstimulated (shaded) or stimulated (unshaded) for 4 h with LPS (100 ng ml⁻¹) or poly(I:C) (10 μg ml⁻¹).

the same genotype yields two types of response to LPS: some cells show a relatively weak response to LPS (low TNF production), and other cells show no response at all (Fig. 6a, bottom). MyD88 deficiency yields a single population of cells that show a uniformly low response to LPS (Fig. 6a, middle) but a dsRNA response that is indistinguishable from the wild type (Fig. 6b, middle), consistent with the results of culture medium assays (Fig. 5c).

Therefore, with regard to the LPS response, macrophages are of two types: Trif-dependent and Trif-independent. Some *Trif*^{Lps2/Lps2} macrophages lack LPS responsiveness, whereas all *Trif*^{Lps2/Lps2} macrophages lack dsRNA responsiveness. Because MyD88 appears to be expressed by all macrophages (Fig. 6a, middle), it would follow that a separate adaptor protein ('adaptor X') may permit the diminished LPS responses that occur in Trif-independent cells. As no TNF is induced by LPS in *Trif*^{Lps2/Lps2};MyD88^{-/-} mutant macrophages (Fig. 5c), we posit that adaptor X is capable of functioning only in cells that co-express MyD88; that is, similar to Mal, adaptor X is MyD88-dependent. Where dsRNA responses are concerned, adaptor X cannot substitute for Trif.

The closest homologue of *Trif* that we have earlier noted to be present in the mouse genome stands as a candidate to fulfil the role of adaptor X, active in the Trif-independent cell population but not in the Trif-dependent cell population. Trif-independent cells are probably capable of specialized responses, by virtue of the fact that they express a unique adaptor molecule.

Receptor diversity meets adaptor degeneracy

We suggest that clinical similarity between certain viral and bacterial diseases is largely explained by the sharing of TIR domain transducers. These transducers serve innate immune responses to microbes that are structurally unrelated, and therefore recognized by different receptors. Trif is one such transducer, and its mutational modification impairs responses to both viral infections and bacterial LPS. Notwithstanding its ability to detect diverse pathogens, the TLR system has a limited repertoire of responses. To some extent, however, degeneracy may be offset by afferent functional specialization among innate immune cells. □

Methods

Innate immune activators, probes and antibodies

Escherichia coli RE595 LPS was a gift of C. Galanos. Synthetic *E. coli* lipid A was produced in the laboratory of S. Kusumoto. dsRNA (poly(I:C)) was purchased from Amersham Pharmacia Biotech. Pam₃CysSer(Lys)₄ was obtained from EMC microcollections GmbH and peptidoglycan was purchased from Fluka. Phosphorothioate-stabilized CpG oligodeoxynucleotide (ODN) 5'-TCCATGACGTCCTGATGCT-3' was obtained from Integrated DNA Technologies. Phosphospecific antibodies for ERKs, p38 and JNK were obtained from New England Biolabs, and antibodies against IκB and IRF-3 were purchased from Santa Cruz Biotechnology. Rabbit anti-mouse iNOS antibodies were obtained from Cayman Chemical. STAT1 [pY701] phosphospecific antibodies and antibodies against total STAT1α were obtained from Biosource International. A MuIL-12p40 ELISA and rMuTNF-α were purchased from Pierce Endogen. rMuIFN-β was obtained from R&D systems. The 340-bp IFN-β hybridization probe was amplified from genomic DNA using a forward primer (5'-CTTCTCCACACAGCCCTCTCCATC-3') and a reverse primer (5'-GAGCAGTTGAGGACATCTCCACGTC-3').

Western blotting

Western analysis of JNK, p38, STAT and ERK1/2 phosphorylation, and degradation of IκB was performed as previously described²⁹. IRF-3 phosphodimer detection was performed as described in ref. 30.

Type I interferon assay

Type I interferon assay was performed according to the method of ref. 31.

Macrophage transfection and FACS analysis

Thioglycolate-elicited peritoneal macrophages were transfected for 48 h with YFP vector alone, or mixed at a 1:3 molar ratio with either *Trif* or *Trif*^{Lps2/Lps2} cDNA expressed under the influence of a CMV promoter. Transfection was performed using Geneporter transfection reagent (Gene Therapy Systems). Cells were induced in the presence of brefeldin A with poly(I:C) (10 μg ml⁻¹), or left uninduced.

Germline mutagenesis

We obtained ENU from Sigma and used it as described¹⁹.

Received 15 March; accepted 7 July 2003; doi:10.1038/nature01889.

Published online 20 July 2003.

- Poltorak, A. *et al.* Defective LPS signaling in C3H/HeJ and C57BL/10ScCr mice: mutations in *Tlr4* gene. *Science* **282**, 2085–2088 (1998).
- Underhill, D. M. *et al.* The Toll-like receptor 2 is recruited to macrophage phagosomes and discriminates between pathogens. *Nature* **401**, 811–815 (1999).
- Takeuchi, O. *et al.* Differential roles of TLR2 and TLR4 in recognition of gram-negative and gram-positive bacterial cell wall components. *Immunity* **11**, 443–451 (1999).
- Hemmi, H. *et al.* A Toll-like receptor recognizes bacterial DNA. *Nature* **408**, 740–745 (2000).
- Alexopoulou, L., Holt, A. C., Medzhitov, R. & Flavell, R. A. Recognition of double-stranded RNA and activation of NF-κB by Toll-like receptor 3. *Nature* **413**, 732–738 (2001).
- Takeuchi, O. *et al.* Discrimination of bacterial lipoproteins by Toll-like receptor 6. *Int. Immunol.* **13**, 933–940 (2001).
- Takeuchi, O. *et al.* Cutting edge: role of Toll-like receptor 1 in mediating immune response to microbial lipoproteins. *J. Immunol.* **169**, 10–14 (2002).
- Beutler, B. Sepsis begins at the interface of pathogen and host. *Biochem. Soc. Trans.* **29**, 853–859 (2001).
- Beutler, B. & Rehli, M. Evolution of the TIR, tolls and TLRs: functional inferences from computational biology. *Curr. Top. Microbiol. Immunol.* **270**, 1–21 (2002).
- Suzuki, N. *et al.* Severe impairment of interleukin-1 and Toll-like receptor signalling in mice lacking IRAK-4. *Nature* **416**, 750–756 (2002).
- Picard, C. *et al.* Pyogenic bacterial infections in humans with IRAK-4 deficiency. *Science* **299**, 2076–2079 (2003).
- Kawai, T., Adachi, O., Ogawa, T., Takeda, K. & Akira, S. Unresponsiveness of MyD88-deficient mice to endotoxin. *Immunity* **11**, 115–122 (1999).
- Kawai, T. *et al.* Lipopolysaccharide stimulates the MyD88-independent pathway and results in activation of IFN-regulatory factor 3 and the expression of a subset of lipopolysaccharide-inducible genes. *J. Immunol.* **167**, 5887–5894 (2001).
- Doyle, S. *et al.* IRF3 mediates a TLR3/TLR4-specific antiviral gene program. *Immunity* **17**, 251 (2002).
- Fitzgerald, K. A. *et al.* Mal (MyD88-adaptor-like) is required for Toll-like receptor-4 signal transduction. *Nature* **413**, 78–83 (2001).
- Horng, T., Barton, G. M. & Medzhitov, R. TIRAP: an adapter molecule in the Toll signaling pathway. *Nature Immunol.* **2**, 835–841 (2001).
- Horng, T., Barton, G. M., Flavell, R. A. & Medzhitov, R. The adaptor molecule TIRAP provides signalling specificity for Toll-like receptors. *Nature* **420**, 329–333 (2002).
- Yamamoto, M. *et al.* Essential role for TIRAP in activation of the signalling cascade shared by TLR2 and TLR4. *Nature* **420**, 324–329 (2002).
- Hoeb, K., Du, X., Goode, J., Mann, N. & Beutler, B. *Lps2*: a new locus required for responses to lipopolysaccharide, revealed by germline mutagenesis and phenotypic screening. *J. Endotoxin Res.* **9**, 250–255 (2003).
- Heppner, G. & Weiss, D. W. High susceptibility of strain A mice to endotoxin and endotoxin-red blood cell mixtures. *J. Bacteriol.* **90**, 696–703 (1965).
- Karaghiosoff, M. *et al.* Central role for type I interferons and Tyk2 in lipopolysaccharide-induced endotoxin shock. *Nature Immunol.* **4**, 471–477 (2003).
- Brown, M. G. *et al.* Vital involvement of a natural killer cell activation receptor in resistance to viral infection. *Science* **292**, 934–937 (2001).
- Smith, H. R. *et al.* Recognition of a virus-encoded ligand by a natural killer cell activation receptor. *Proc. Natl Acad. Sci. USA* **99**, 8826–8831 (2002).
- Yamamoto, M. *et al.* Cutting edge: a novel Toll/IL-1 receptor domain-containing adapter that preferentially activates the IFN-β promoter in the Toll-like receptor signaling. *J. Immunol.* **169**, 6668–6672 (2002).
- Oshiumi, H., Matsumoto, M., Funami, K., Akazawa, T. & Seya, T. TICAM-1, an adaptor molecule that participates in Toll-like receptor 3-mediated interferon-beta induction. *Nature Immunol.* **4**, 161–171 (2003).
- Xu, Y. *et al.* Structural basis for signal transduction by the Toll/interleukin-1 receptor domains. *Nature* **408**, 111–115 (2000).
- Du, X., Poltorak, A., Silva, M. & Beutler, B. Analysis of Tlr4-mediated LPS signal transduction in macrophages by mutational modification of the receptor. *Blood Cells Mol. Dis.* **25**, 328–338 (1999).
- Fitzgerald, K. A. *et al.* IKKε and TBK1 are essential components of the IRF3 signaling pathway. *Nature Immunol.* **4**, 491–496 (2003).
- Kim, S. O., Ono, K. & Han, J. Apoptosis by pan-caspase inhibitors in lipopolysaccharide-activated macrophages. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **281**, L1095–L1105 (2001).
- Iwamura, T. *et al.* Induction of IRF-3/7 kinase and NF-κB in response to double-stranded RNA and virus infection: common and unique pathways. *Genes Cells* **6**, 375–388 (2001).
- Orange, J. S. & Biron, C. A. Characterization of early IL-12, IFN-αβ, and TNF effects on antiviral state and NK cell responses during murine cytomegalovirus infection. *J. Immunol.* **156**, 4746–4756 (1996).
- Poltorak, A. *et al.* Genetic and physical mapping of the *Lps* locus—identification of the toll-4 receptor as a candidate gene in the critical region. *Blood Cells Mol. Dis.* **24**, 340–355 (1998).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We are grateful to S. Akira for providing MyD88-knockout mice for use in these studies; S. Kusumoto for providing synthetic lipid A; C. Galanos for providing highly purified LPS; and O. Haller for discussions. This work was supported by funding from NIH grants and by a grant from DARPA.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.B. (bruce@scripps.edu).

A γ -ray burst with a high-energy spectral component inconsistent with the synchrotron shock model

M. M. González^{1,2}, B. L. Dingus^{1,2}, Y. Kaneko³, R. D. Preece³, C. D. Dermer⁴ & M. S. Briggs³

¹Physics Department, University of Wisconsin, Madison, Wisconsin 53706, USA

²Los Alamos National Laboratory, MS H803, Los Alamos, New Mexico 87545, USA

³Physics Department, University of Alabama in Huntsville, National Space Science and Technology Center, Huntsville, Alabama 35899, USA

⁴Code 7653, Naval Research Laboratory, Washington DC 20375, USA

Gamma-ray bursts are among the most powerful events in nature. These events release most of their energy as photons with energies in the range from 30 keV to a few MeV, with a smaller fraction of the energy radiated in radio, optical, and soft X-ray afterglows¹. The data are in general agreement with a relativistic shock model², where the prompt and afterglow emissions³ correspond to synchrotron radiation from shock-accelerated electrons. Here we report an observation of a high-energy (multi-MeV) spectral component in the burst of 17 October 1994 that is distinct from the previously observed lower-energy γ -ray component. The flux of the high-energy component decays more slowly and its fluence is greater than the lower-energy component; it is described by a power law of differential photon number index approximately -1 up to about 200 MeV. This observation is difficult to explain with the standard synchrotron shock model², suggesting the presence of new phenomena such as a different non-thermal electron process, or the interaction of relativistic protons with photons at the source.

Prior EGRET observations⁴ of photons with energies >100 MeV from four gamma-ray bursts (GRBs) display spectra consistent with an extension of the electron synchrotron component. In the case of GRB940217, GeV photons were detected in the prompt phase and in an extended emission phase that lasted for over 90 min after the start of the GRB⁵. Milagrito observations⁶ of GRB970417 at energies near ~ 0.1 TeV hinted at a distinct higher-energy component, but lacked energy resolution to provide a spectrum. A related self-Compton emission component, due to the relativistic electrons scattering their own synchrotron radiation, is expected^{7,8} at GeV–TeV energies. This component, or emission from interactions of the blast wave particles with a surrounding cloud⁹, could explain the highest-energy γ -rays detected from GRB940217 and GRB970417.

Finally, two hard-spectrum events that could be GRBs have been previously detected¹⁰ with the Solar Maximum Mission in a restricted energy range from 0.8 to 10 MeV, but these observations lacked broadband spectral coverage in the regime between tens of keV to hundreds of MeV to relate or distinguish it from the synchrotron component. Thus previous observations of high-energy emission from GRBs are consistent with the relativistic shock model.

We have combined and analysed data from two detectors on the Compton Observatory, one of the BATSE's Large Area Detectors¹¹, LAD, and EGRET's calorimeter TASC¹² (Total Absorption Shower Counter), to obtain prompt GRB spectra over a broad energy range extending from 30 keV to 200 MeV. TASC is generally sensitive to the brightest bursts observed with BATSE. In 26 BATSE bursts that were bright at 300 keV and were also detected with TASC, the high-energy spectra are consistent with the single spectral component observed by BATSE in all cases except for the GRB of 17 October 1994 (GRB941017).

GRB941017 occurred at 10:19:33.938 UT in the galactic-coordinate direction $l = 50.5^\circ$ and $b = -11.7^\circ$, and was the eleventh highest fluence burst observed by BATSE in its 9-yr lifetime. Ninety per cent of the flux observed with BATSE occurred in a time interval of 77 s. Properties of the prompt emission were previously reported for BATSE^{13,14}, COMPTEL¹⁵ and TASC¹⁶ observations in the energy ranges of 30 keV–2 MeV, 700 keV–30 MeV and 1–200 MeV, respectively. The burst direction was 66° with respect to the pointing-axis direction of the Compton Observatory, reducing COMPTEL's effective area and placing the burst outside the ~ 1 sr field of view of the EGRET spark chamber, which was sensitive above 30 MeV.

The light curves of the emission as observed by the LAD and the TASC are shown in Fig. 1. The TASC detected seven 32.768-s spectra that were more than three standard deviations above the background. The TASC background was obtained¹⁷ by fitting a polynomial to time intervals on each side of the burst, and was further checked against a time interval 15 orbits earlier when the spacecraft was located at the same geomagnetic rigidity. The spectrum accumulation begins 18 s before the BATSE trigger. Background-subtracted spectra of LAD and TASC data were fitted jointly using the two detectors' separate response matrices. The LAD data were binned in time so as to match the 32.768-s time resolution of the TASC data.

We analysed the data set using the empirical Band's GRB function¹⁸ that consists of two smoothly joined power laws. This function is characterized by the amplitude A , and power-law photon number indices α (β) below (above) a break energy, which is uniquely related to the energy E_{peak} with the maximum

Table 1 Spectral fitting parameters of the differential photon flux

Time from BATSE trigger (s)	–18 to 14	14 to 47	47 to 80	80 to 113	113 to 211
Band GRB function parameters					
A (photons $\text{cm}^{-2} \text{s}^{-1} \text{keV}^{-1}$)	0.01 ± 0.0005	0.06 ± 0.001	0.02 ± 0.001	0.007 ± 0.003	$0.05^{+0.04}_{-0.02}$
E_{peak} (keV)	505 ± 69	350 ± 8	240 ± 14	91 ± 15	10 fixed
α (low-energy spectral index)	-0.84 ± 0.09	-0.79 ± 0.02	-1.08 ± 0.06	-1.45 ± 0.27	-1.00 fixed
β (high-energy spectral index)	-2.22 ± 0.13	-2.46 ± 0.05	$-2.65^{+0.22}_{-0.88}$	$-3.13^{+0.63}_{-\infty}$	-2.73 ± 0.55
High-energy power law parameters					
A_{PL} ($\times 10^{-7}$ photons $\text{cm}^{-2} \text{s}^{-1} \text{keV}^{-1}$)	NA	2.4 ± 0.6	2.4 ± 1.0	3.2 ± 0.7	1.8 ± 0.5
γ (highest-energy spectral index)	NA	-1.00 fixed	$-1.06^{+0.70}_{-0.44}$	$-1.10^{+0.32}_{-0.17}$	$0.96^{+0.36}_{-0.17}$
Probability	NA	1.40×10^{-4}	3.71×10^{-4}	6.49×10^{-8}	1.44×10^{-7}
Energy flux for different energy ranges ($\times 10^{-6} \text{ergs}^{-1} \text{cm}^{-2}$)					
0.03–2 MeV	0.85 ± 0.02	3.52 ± 0.46	0.96 ± 0.19	0.20 ± 0.02	0.03 ± 0.004
2–10 MeV	0.33 ± 0.008	0.82 ± 0.11	0.20 ± 0.04	0.15 ± 0.01	0.07 ± 0.007
10–200 MeV	0.38 ± 0.009	2.63 ± 0.35	2.03 ± 0.40	2.54 ± 0.25	1.67 ± 0.18

The first six rows contain the best spectral fit parameters for the five time intervals shown in Fig. 2. The seventh row shows the probability that the improvement in χ^2 from the addition of the high-energy power law in the fit is due to chance, as determined by the χ^2 test. The last three rows contain the energy fluxes for three different energy ranges. For the first interval, the high-energy power law was not required to fit the spectrum. For the second and last intervals, poorly determined parameters were fixed at values that are either consistent with those determined for later time intervals (γ in 14–47 s) or assumed to be reasonable (E_{peak} and α in 113–211 s). The temporal fit of the energy fluxes (F) in the last four time intervals when described by $F = At^{-\phi}$, yields $\phi = 2.8, 1.45$ and 0.2 for the energy ranges of 0.03–2, 2–10 and 10–200 MeV, respectively. Thus the temporal energy flux decay of the higher-energy component is much slower than the typical energy flux decay of afterglows. NA, not applicable.

energy flux. An additional power law function, described by the expression $A_{\text{PL}}[E(\text{keV})/30 \text{ MeV}]^\gamma$, was required to fit the higher-energy γ -ray excess at later times. In addition, a calibration factor of 0.45 was applied to the TASC data to normalize the flux relative to LAD. This factor is similar to that required in other joint BATSE–TASC fits of GRBs¹⁷, is consistent with the errors in the calculated effective area as determined from the pre-flight calibration¹⁹, and has little effect on the parameter values of the fit.

The data and spectral fits for five time intervals are shown in Fig. 2, with the best-fit parameters given in Table 1. The last time interval shown is the summation of three 32.768-s intervals. The parameters of the Band GRB function and the temporal evolution of the lower-energy emission component are consistent with many other bursts observed by BATSE²⁰. For example, E_{peak} generally decreases with time and flux, in accord with the well-known hard-to-soft evolution in GRB spectra²¹. The higher-energy component, which is represented by a power law, remains bright and the spectral index does not change within the statistical uncertainties. At the highest energies, the flux remains constant throughout the GRB within statistical uncertainties, while the flux at 300 keV decays by

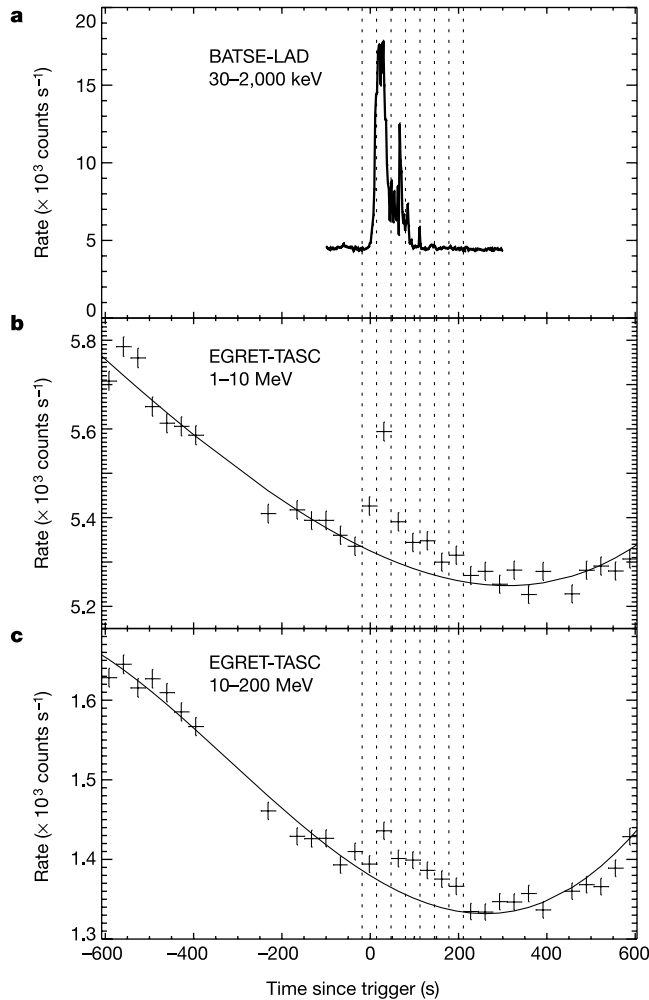


Figure 1 Count rates for GRB941017. **a**, From LAD; **b**, **c**, from TASC. The LAD data show a weak precursor ~ 90 s before the trigger, which is not included in the background estimation. The TASC background fit is shown as a line in **b** and **c**. In the time interval from 113 to 211 s, LAD flux is also still significantly detected with a 17σ excess over the background while TASC flux in the energy range of 1–200 MeV is only 6σ above the background. However, the TASC flux corresponds to a much larger fraction of the TASC detected emission than is seen in the LAD data. As shown in **c**, the TASC flux is also 6σ above the background in the energy range of 10–200 MeV in this same time interval.

about three orders of magnitude. The difference in the temporal evolution of the low- and high-energy components can also be seen in Table 1, which shows the energy flux in three different energy intervals: 30 keV–2 MeV, 2 MeV–10 MeV and 10 MeV–200 MeV for the five time intervals. The additional high-energy component, with its longer duration, results in a total >30 keV fluence of $6.5 \times 10^{-4} \text{ erg cm}^{-2}$, which is more than three times that estimated from the BATSE energy range alone. The high-energy power law does not exhibit a cut-off, suggesting that even more energy is emitted above 200 MeV. The extension of the higher-energy component to lower energies is also apparent in the LAD data at late times.

GRBs are expected to exhibit a self-Compton component during the prompt phase that extends to photon energies $> \gamma_{\text{pk}}^2 E_{\text{peak}}$, where γ_{pk} is the electron Lorentz factor producing synchrotron emission at E_{peak} . Because $\gamma_{\text{pk}} \gg 10^3$ in the external shock of the relativistic

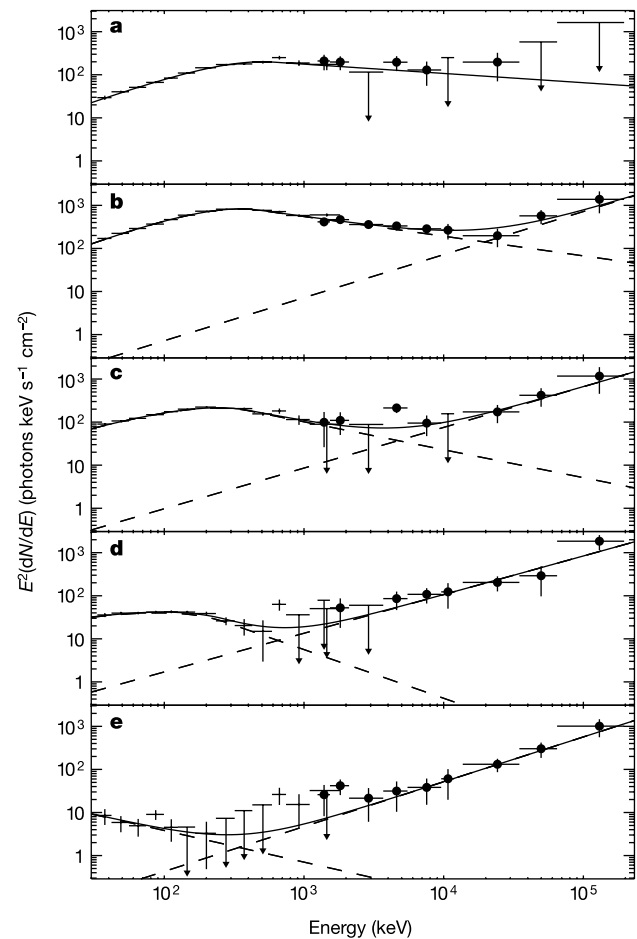


Figure 2 Energy fluxes from GRB941017. **a–e**, Data were obtained with LAD (crosses) and TASC (filled circles) during five time intervals shown in Table 1 (with **a** and **e** being the earliest and latest time intervals, respectively), which covers the period of significant detection with LAD and TASC and shows the synchrotron and the new high-energy components. For the purpose of the plot, but not for the spectral fit, the TASC data are binned in energy to give at least 2σ significance over background. Solid curves show model fits to the data using the parameters given in Table 1 and the spectral model described in the text. Upper limits are 2σ . The two spectral components—the Band GRB function at lower energies and the higher-energy power law (dashed lines)—are most obvious at later times, but both components could be present in all the time intervals. Both detectors independently observed the high-energy component at later times. Furthermore, the fit to only TASC data is in agreement with the joint fit. The low-energy (<10 MeV) emission typically associated with synchrotron radiation decays more rapidly than the 10–200 MeV emission.

blast wave², this places the peak of the self-Compton component at GeV–TeV energies, even taking into account Klein–Nishina effects on Compton scattering^{7,8}. To produce self-Compton fluxes with brightness comparable to the synchrotron flux near E_{peak} , as is necessary to explain observations of GRB941017, would require orders-of-magnitude greater luminosities in the self-Compton component, making such an interpretation for the high-energy emission untenable. A related possibility is that the enhanced MeV emission in GRB941017 is due to Compton radiation from synchrotron-emitting electrons that scatter photons from the internal shocks thought to produce the prompt radiation²², or from an intense radiation field that originates from outside the relativistic blast wave shell²³. The evolution of the Compton fluxes should, however, follow the behaviour of the synchrotron emission, contrary to observations that the high-energy component evolves with time independently of the synchrotron component. This behaviour could also pose difficulties for other electron models involving comptonization of photospheric and reverse shock radiation (see ref. 22 for a summary).

Another explanation for the observations is that ultra-relativistic hadrons produce the multi-MeV γ -rays by inducing electromagnetic cascades through photomeson and photo-pair production. Acceleration of high-energy protons by GRBs is proposed in models that link GRBs with the sources of the highest-energy cosmic rays^{24,25}, and would be confirmed through detection of high-energy neutrinos²⁶. Hadronic cascade radiation spectra are hard and will decay more slowly than the lepton radiations²⁷, with potentially brighter fluxes in the hadronic component if more energy is found in the hadrons than the electrons and the hadronic radiation efficiency is high.

Detailed theoretical modelling of this observation is required to quantify these constraints on the production mechanisms. Unfortunately, this observation does not extend to high enough γ -ray energies to know the total luminosity in the high-energy component or the γ -ray energy at which the flux peaks. However, future observations by GLAST²⁸ (a γ -ray observatory scheduled to be launched in 2006), with its Gamma Ray Burst Monitor (10 keV–25 MeV) and its Large Area Telescope (20 MeV–300 GeV), will be more sensitive and extend to higher energies, which will further clarify this possible GRB/cosmic ray connection.

Note added in proof: After the completion of the analysis done in this letter, the existence of COMPTEL BSA data (300 keV–10.6 MeV) covering portions of GRB941017 was brought to our attention. Subsequent analysis shows that these data are consistent with the results herein. (L. Hanlon, personal communication). □

Received 23 April; accepted 27 June 2003; doi:10.1038/nature01869.

1. van Paradijs, J., Kouveliotou, C. & Wijers, R. A. M. J. Gamma-ray burst afterglows. *Annu. Rev. Astron. Astrophys.* **38**, 379–425 (2000).
2. Mészáros, P. Theories of gamma-ray bursts. *Annu. Rev. Astron. Astrophys.* **40**, 137–169 (2002).
3. Costa, E. *et al.* Discovery of an X-ray afterglow associated with the gamma-ray burst of 28 February 1997. *Nature* **387**, 783–785 (1997).
4. Dingus, B. L. in *High-Energy Gamma Ray Astronomy* (eds Aharonian, F. A. & Völk, H. J.) 383–391 (AIP, New York, 2001).
5. Hurley, K. *et al.* Detection of a gamma-ray burst of very long duration and very high energy. *Nature* **372**, 652–654 (1994).
6. Atkins, R. *et al.* Evidence for TeV emission from GRB 970417A. *Astrophys. J.* **533**, L119–L122 (2000).
7. Dermer, C. D., Chiang, J. & Mitman, K. E. Beaming, baryon loading, and the synchrotron self-compton component in gamma-ray bursts. *Astrophys. J.* **537**, 785–795 (2000).
8. Zhang, B. & Mészáros, P. High-energy spectral components in gamma-ray burst afterglows. *Astrophys. J.* **559**, 110–122 (2001).
9. Katz, J. I. Delayed hard photons from gamma-ray bursts. *Astrophys. J.* **432**, L27–L29 (1994).
10. Harris, M. J. & Share, G. H. A Search for hard-spectrum gamma-ray bursts using SMM. *Astrophys. J.* **494**, 724–733 (1998).
11. Paciesas, W. S. *et al.* The fourth BATSE gamma-ray burst catalog (revised). *Astrophys. J. Suppl.* **122**, 465–495 (1999).
12. Esposito, J. A. *et al.* In-flight calibration of EGRET on the Compton gamma-ray observatory. *Astrophys. J. Suppl.* **123**, 203–217 (1999).
13. Preece, R. D. *et al.* The BATSE gamma-ray burst spectral catalog. I. High time resolution spectroscopy of bright bursts using high energy resolution data. *Astrophys. J. Suppl.* **126**, 19–36 (2000).
14. Briggs, M. S. *et al.* in *Gamma-ray Bursts* (eds Meegan, C. A., Preece, R. D. & Koshut, T. M.) 299–303 (AIP, New York, 1998).

15. COMPTEL burst images and time profiles. (<http://wwwgro.unh.edu/bursts/cgrbdata.html>); modified 13 August 1997.
16. Catelli, J. R. & Dingus, B. L. in *Gamma-ray Bursts* (eds Meegan, C. A., Preece, R. D. & Koshut, T. M.) 309–313 (AIP, New York, 1998).
17. González, M. M. *et al.* in *Gamma-Ray Burst and Afterglow Astronomy 2001* (eds Ricker, G. R. & Vanderspek, R. K.) 267–269 (AIP, New York, 2003).
18. Band, D. *et al.* BATSE observations of gamma-ray burst spectra. I — Spectral diversity. *Astrophys. J.* **413**, 281–292 (1993).
19. Thompson, D. J. *et al.* Calibration of the Energetic Gamma-Ray Experiment Telescope (EGRET) for the Compton Gamma-Ray Observatory. *Astrophys. J. Suppl.* **86**, 629–656 (1993).
20. Liang, E. & Kargatis, V. Dependence of the spectral evolution of gamma-ray bursts on their photon fluence. *Nature* **381**, 49–51 (1996).
21. Fishman, G. J. & Meegan, C. A. Gamma-ray bursts. *Annu. Rev. Astron. Astrophys.* **33**, 415–458 (1995).
22. Zhang, B. & Mészáros, P. An analysis of gamma-ray spectral break models. *Astrophys. J.* **581**, 1236–1247 (2002).
23. Inoue, S., Guetta, D. & Pacini, F. Precursor plerionic activity and high-energy gamma-ray emission in the supernova model of gamma-ray bursts. *Astrophys. J.* **583**, 379–390 (2003).
24. Vietri, M. The acceleration of ultra-high-energy cosmic rays in gamma-ray bursts. *Astrophys. J.* **453**, 883–889 (1995).
25. Waxman, E. Cosmological gamma-ray bursts and the highest energy cosmic rays. *Phys. Rev. Lett.* **75**, 386–389 (1995).
26. Alvarez-Muñiz, J., Halzen, F. & Hooper, D. W. High energy neutrinos from gamma ray bursts: Event rates in neutrino telescopes. *Phys. Rev. D.* **62**, 093015 (2000).
27. Boettcher, M. & Dermer, C. High-energy gamma rays from ultra-high-energy cosmic-ray protons in gamma-ray bursts. *Astrophys. J.* **499**, L131–L134 (1998).
28. GLAST. (<http://glast.gsfc.nasa.gov/>); modified 29 April 2003.

Acknowledgements We thank M. Harris for discussions, and D. L. Bertsch for assistance with accessing the EGRET data.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.M.G. (magda@whopper.lanl.gov).

Erased electrostatic lithography for quantum components

Rolf Crook, Abi C. Graham, Charles G. Smith, Ian Farrer, Harvey E. Beere & David A. Ritchie

Department of Physics, University of Cambridge, Madingley Road, Cambridge CB3 0HE, UK

Quantum electronic components^{1,2}—such as quantum antidots and one-dimensional channels—are usually defined from doped GaAs/AlGaAs heterostructures using electron-beam lithography or local oxidation by conductive atomic force microscopy^{3,4}. In both cases, lithography and measurement are performed in very different environments, so fabrication and test cycles can take several weeks. Here we describe a different lithographic technique, which we call erased electrostatic lithography (EEL), where patterns of charge are drawn on the device surface with a negatively biased scanning probe in the same low-temperature high-vacuum environment used for measurement. The charge patterns locally deplete electrons from a subsurface two-dimensional electron system (2DES) to define working quantum components. Charge patterns are erased locally with the scanning probe biased positive or globally by illuminating the device with red light. We demonstrate and investigate EEL by drawing and erasing quantum antidots, then develop the technique to draw and tune high-quality one-dimensional channels^{5,6}. The quantum components are imaged using scanned gate microscopy^{7–11}. A technique similar to EEL has been reported previously, where tip-induced charging of the surface or donor layer was used to locally perturb a 2DES before charge accumulation imaging¹².

A long quantum wire, of lithographic length $5\text{ }\mu\text{m}$ and width $1\text{ }\mu\text{m}$, is used to characterize the quantum components drawn by EEL. The wire is defined from a 2DES, with electron mobility $5.0 \times 10^6\text{ cm}^2\text{ V}^{-1}\text{ s}^{-1}$ and density $3.1 \times 10^{11}\text{ cm}^{-2}$, formed at a GaAs/AlGaAs heterojunction 97 nm beneath the GaAs surface. A gate bias V_g is applied to metal surface electrodes, which are fabricated using electron-beam technology and are visible in the topographic image of Fig. 1a. When V_g is sufficiently negative to deplete electrons from the underlying 2DES, a regime known as 'definition', the quantum wire forms between the electrodes. The width, and therefore the conductance G , of the wire are reduced by making V_g more negative until the device is said to be 'pinched off' when G reaches zero. The scanning probe is a conductive atomic force microscope tip biased to V_p (ref. 13). During experiments, the sample and scanning probe are cooled to 20 mK in a dilution refrigerator.

The first set of experiments investigate the EEL mechanism using

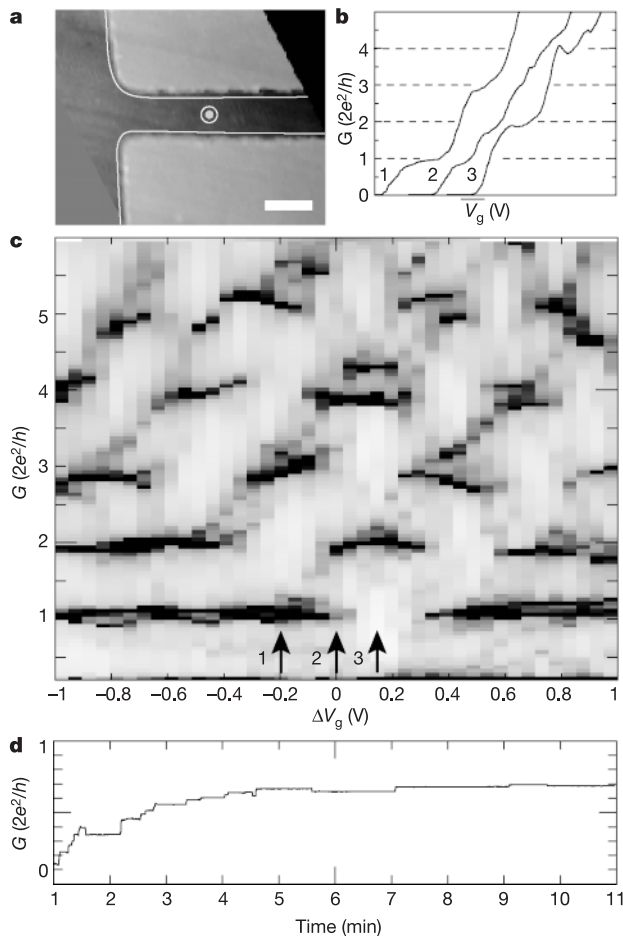


Figure 1 Characterization of an EEL charge spot defining a quantum antidot in the 2DES. **a**, Topographic image locating the wire surface electrodes and a charge spot drawn with $V_p = -6\text{ V}$. The electron depletion is outlined in white. (Scale bar, $1\text{ }\mu\text{m}$.) **b**, Plots of G as a function of V_g for three different ΔV_g identified by arrows in **c**. Plateaus are seen at certain integers of $2e^2/h$. **c**, Greyscale plots of the differential of conductance dG/dV_g . For each combination of V_g and ΔV_g , dG/dV_g is calculated to determine the colour of the pixel at position $(\Delta V_g, G)$. Black regions correspond to plateaus in wire conductance and white regions correspond to the transitions between plateaus. Regions where dG/dV_g is multivalued in G are interpreted as plateaus and coloured black. **d**, Conductance of the wire as a function of time after a charge spot was drawn with $V_p = -3\text{ V}$. The wire conductance was initially set to $3 \times 2e^2/h$ by adjusting V_g . A $1.1\text{ k}\Omega$ series resistance has been subtracted.

a single spot of charge drawn near the lateral centre of the wire with $V_p = -6\text{ V}$. This creates a potential hill, or quantum antidot, in the 2DES (Fig. 1a) which defines two short and narrow conductive paths^{14,15}, called point contacts, connected in parallel within the width of the wire. Electron transport is ballistic and the parallel point contacts form the narrowest part of the wire, so G is determined by the sum of the point contacts' conductances. By applying different biases to the two surface electrodes, the widths of the two point contacts are controlled independently, where $\overline{V_g}$ is the average electrode bias and the difference between the electrode biases is the offset bias ΔV_g . Changing ΔV_g can also be viewed as shifting the quantum wire laterally while the antidot remains stationary¹⁴. When plotted as a function of width or electron density, the conductance of a point contact is quantized in integers of $2e^2/h$, indicating the presence of a one-dimensional (1D) electron system^{5,6}.

The antidot is studied by recording G as a function of $\overline{V_g}$ for different ΔV_g . Figure 1b presents three conductance plots with ΔV_g identified by the corresponding arrow in Fig. 1c which plots the differential of conductance $dG/d\overline{V_g}$ against G and ΔV_g as a grey-scale where conductance plateaus are black. When ΔV_g is adjusted so that the antidot is at the exact lateral centre of the wire, the point contacts are the same width and their conductance plateaus add to produce plateaus in G at only even integers of $2e^2/h$. This occurs when $\Delta V_g = 0.15$ (see arrow and plot labelled 3 in Fig. 1), showing that the charge spot was not drawn at the exact lateral centre. At $\Delta V_g = 0\text{ V}$ and $\Delta V_g = 0.3\text{ V}$ the point contacts alternately contribute plateau to the wire conductance and plateaus are seen at all integers of $2e^2/h$. At $\Delta V_g = -0.2\text{ V}$ and $\Delta V_g = 0.4\text{ V}$ adjacent plateaus from the point contacts add to generate plateaus at only odd integers of $2e^2/h$. This odd-even pattern continues as ΔV_g is taken to more positive and negative voltages. When the wire is defined, a maximum of four plateaus are seen from each point contact. Each plateau corresponds to the population of a 1D sub-band^{1,2}, and each sub-band adds half an electron wavelength to the channel width. With four sub-bands occupied, a typical electron

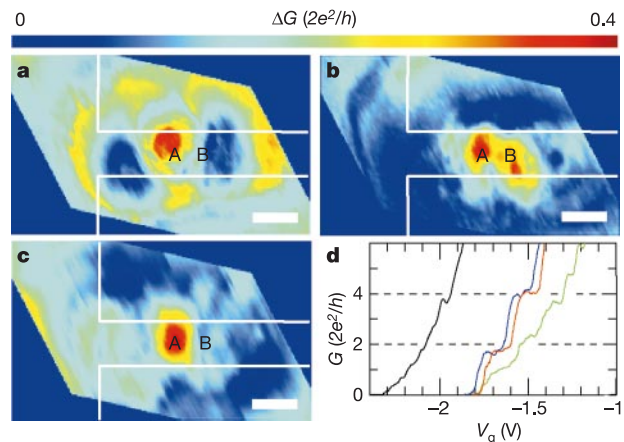


Figure 2 SGM images of an EEL sequence. The images are made by raster scanning the probe 50 nm above the device surface with $V_p = -3\text{ V}$. The location of the wire surface electrodes are outlined on the images. **a**, SGM image made after a spot of charge with $V_p = -6\text{ V}$ was drawn at point A. **b**, SGM image made after a second spot of charge was drawn at point B, which is 750 nm from point A. **c**, SGM image made after the second spot was erased by returning to point B with $V_p = +3\text{ V}$. Structure remote from points A and B is due to disorder from the wire. **d**, Conductance plots made during the EEL sequence. The black plot was made before any patterning, the red plot after the first spot was patterned, the green plot after the second spot, and the blue plot after the second spot was erased. The blue and red plots are quantized in even integers of $2e^2/h$, because the charge spot created an antidot central in the wire. A $1.1\text{ k}\Omega$ series resistance has been subtracted. Scale bars, $1\text{ }\mu\text{m}$.

wavelength is 120 nm (refs 16, 17), so with a defined wire width of 800 nm the antidot radius is $(800 \text{ nm} - 2 \times 4 \times 60 \text{ nm})/2 = 160 \text{ nm}$.

To investigate the time evolution of EEL charge, Fig. 1d plots G as a function of time after a charge spot was drawn near the wire centre. During the first 5 minutes the conductance increases in small steps owing to single electrons moving to reduce the local peak in surface-charge density. After this initial charge relaxation, charge drawn with a negative V_p remained unchanged for the duration of the experiment, which in this case was about one week. Charge from spots drawn with a positive V_p leak away within an hour, so a positive bias is used to locally erase patterns drawn with a negative bias. Charge patterns are globally erased by illuminating with a red light-emitting diode for a few seconds. Charge spots drawn with $V_p < -5 \text{ V}$ locally deplete electrons from the 2DES, and spots drawn with $-5 \text{ V} < V_p < 0 \text{ V}$ locally reduce the electron density. These leakage and bias parameters are similar to those found using surface electrodes, so we propose that the EEL mechanism is the charging of GaAs surface states. The number of electrons n in a charge spot drawn with $V_p = -6 \text{ V}$ is approximated by multiplying the antidot area by the 2DES electron density giving 240 electrons, and the capacitance from the charge spot to the 2DES is $C = -ne/V_p = 6 \text{ aF}$. By assuming that 2DES lateral depletion is equal to the 2DES depth, the radius of the charge spot is approximately 60 nm and, because charge rearrangement is observed, the maximum charge spot electron density is $2 \times 10^{12} \text{ cm}^{-2}$.

Scanned gate microscopy (SGM) generates images by scanning a biased probe over a quantum wire, while the wire conductance is recorded to determine the pixel colour^{7–11}. SGM images highlight regions that determine the wire conductance, and can therefore be used to locate certain EEL defined quantum components and conveniently requires no modification to the apparatus. Whether the wire conductance increases or decreases as the probe passes depends upon the SGM probe bias, the EEL probe bias, and the type of EEL component. Figure 2 presents a series of SGM images made

between EEL charge spot operations. Single (Fig. 2a) and double (Fig. 2b) peaks are clearly resolved in the SGM images after one and two spots, respectively, are drawn. The second spot of charge is locally erased, and the SGM image again shows a single peak (Fig. 2c). SGM features often appear circular because the conductance is determined by the capacitance between the probe and the component which is a predominantly radial function^{8,9}. Figure 2d shows the corresponding plots of G against V_g .

The second set of experiments develop EEL to fabricate high-quality point contacts. Two EEL charge spots with $V_p = -5 \text{ V}$ were drawn offset across the width of the wire, as identified by points A and C in the upper inset of Fig. 3, to define a point contact within the wire. Figure 3 plots G against V_g , and plateaus in G are seen to be quantized in units of $2e^2/h$, demonstrating that a 1D electron system has formed. Although the plateaus are numerous, structure on the plateaus suggests that disorder from the quantum wire is limiting the quality of this device. The quantum wire is known to be disordered, because plateaus are not seen before charge is drawn. A series of conductance plots, including one before charge is drawn, are shown in the lower inset of Fig. 3. A single spot of charge is drawn at point B and poorly defined plateaus are seen. After the charge spot is locally erased, the conductance plot is almost identical to the first. This inset also shows the creation of a barrier by drawing charge spots at points A, B and C, which are separated by 300 nm.

A higher-quality EEL point contact was fabricated by locating it outside the disordered quantum wire. Linear barriers were defined from the 2DES by drawing lines of spots of charge, the spots being separated by 100 nm, as illustrated in the upper inset of Fig. 4. Figure 4 plots point contact conductance as a function of V_p (not V_g). Crucially, the probe is now positioned 50 nm above the device surface, so EEL charge patterns are preserved despite the large V_p . V_g is kept constant, and functions only to isolate the source and drain connections. Comparing Fig. 4 with Fig. 3, the range of V_p is much larger than the range of V_g . This is because the gate to point contact capacitance is smaller when the gate is the scanning probe rather than the surface electrodes, owing to the probe geometry and screening from surface states^{8,9}. With the point contact located outside the wire, the plateaus observed in G are better defined and

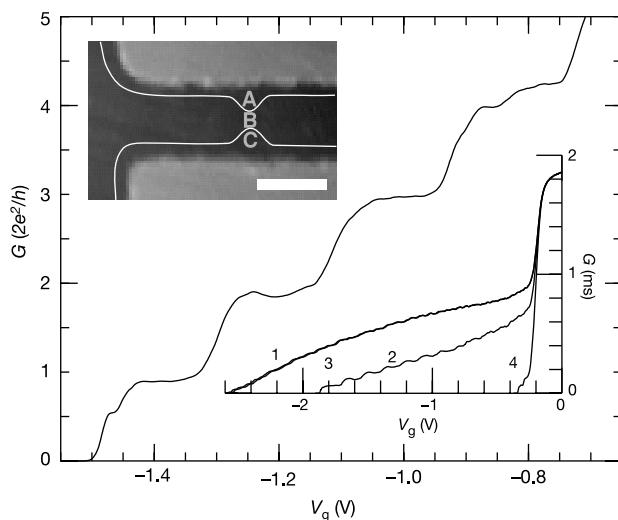


Figure 3 Characterization of an EEL point contact located inside the quantum wire. The upper inset shows a topographic image where the white line outlines electron depletion after spots of charge were drawn at points A and C with $V_p = -5 \text{ V}$. The main figure plots G against V_g , and plateaus are seen at integers of $2e^2/h$. The lower inset plots G against V_g during a different EEL sequence. Plot 1 was taken before any charge spots, and plot 2 after a spot of charge was drawn at B with $V_p = -6 \text{ V}$. Plateaus roughly quantized in integers of $2e^2/h$ suggest that an antidot, offset from the centre, has formed. Plot 3 was taken after the spot at B was erased with $V_p = +3 \text{ V}$ and largely overlaps plot 1. Plot 4 shows the wire characteristics after spots of charge were drawn at A, B and C with $V_p = -6 \text{ V}$ to complete a barrier which pinches off as soon as the wire is defined. A $1.1 \text{ k}\Omega$ series resistance has been subtracted from the main figure. Scale bar, $1 \mu\text{m}$.

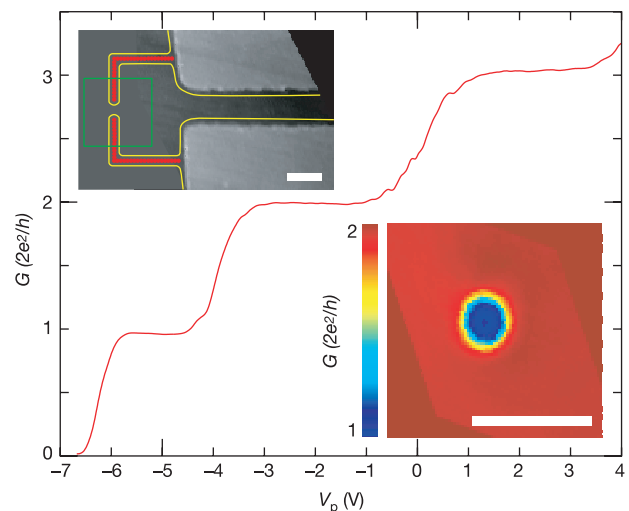


Figure 4 Characterization of an EEL point contact located outside the quantum wire. The upper inset shows a topographic image where the yellow line outlines electron depletion after spots of charge were drawn at all the red points with $V_p = -6 \text{ V}$. With the probe centred over the point contact and 50 nm off the surface, the main figure plots G against V_p and plateaus are observed at integers of $2e^2/h$. The lower inset presents an SGM image made with $V_p = -4.5 \text{ V}$ of the area outlined in the upper inset. A $1.1 \text{ k}\Omega$ series resistance has been subtracted. Scale bars, $1 \mu\text{m}$.

extremely flat, meaning disorder is reduced considerably. The conductance deviates by less than 1% over the plateaus, showing that EEL point contacts are of comparable or higher quality than those fabricated by other techniques^{4,18,19}. Given that the EEL point contact is short, the width of the plateaus compared to the width of the transitions indicates that lateral confinement is strong²⁰. The lower inset of Fig. 4 presents an SGM image to precisely locate the point contact. A unique feature of EEL is the ability to tune a quantum component. In the case of the point contact, additional spots of charge, separated by 10 nm, reliably shift the conductance plots by 1/30 of a plateau period. With the point contact located outside the wire, EEL tuning provides control over geometry while V_p independently controls electron density. Independent control of point contact geometry and electron density is desired to study electron interaction effects in 1D electron systems¹⁹. We note that a feature near $0.7 \times 2e^2/h$ is often seen when the point contact is drawn inside the wire, but is not seen when the point contact is drawn outside, which may be due to reduced electron density or point contact geometry. An EEL point contact tuned to transmit less than one sub-band has an immediate application as a low-noise charge detector for quantum systems²¹.

The device structure is drawn and modified during EEL. The present charge spot area is an order of magnitude larger than the probe to surface contact area²², so smaller charge spots, and therefore higher EEL resolution, would be attained using a less negative probe bias with a shallower 2DES. A shallow 2DES would further improve resolution, as 2DES lateral depletion would decrease²³. Lithography and test cycles take several weeks using conventional electron-beam or atomic force microscopy lithography, whereas EEL-drawn components take only a few hours to optimize. EEL is a particularly productive technique where the device geometry is of interest, such as the study of chaotic electron trajectories in large open dots²⁴, or interaction effects seen in 1D transport¹⁹. EEL may prove to be particularly useful in the construction of a solid-state scaleable quantum computer, where the required level of uniformity between quantum components is hard to achieve using other schemes²⁵. A multifunction scanning probe would visit each component in turn to characterize it, to repair it if necessary, and then to tune it using EEL, so producing the desired array of electrically identical components. A further advantage is that EEL can distance quantum components from surface electrodes, thereby reducing electrical noise and increasing phase coherence times. □

Received 6 March; accepted 18 June 2003; doi:10.1038/nature01841.

1. Beenakker, C. W. J. & van Houten, H. in *Solid State Physics* Vol. 44 (eds Ehrenreich, H. & Turnbull, D.) 1–228 (Academic, New York, 1991).
2. Smith, C. G. Low-dimensional quantum devices. *Rep. Prog. Phys.* **59**, 235–282 (1996).
3. Held, R., Heinzl, T., Studerus, P., Ensslin, K. & Holland, M. Semiconductor quantum point contact fabricated by lithography with an atomic force microscope. *Appl. Phys. Lett.* **71**, 2689–2691 (1997).
4. Curson, N. J. *et al.* Ballistic transport in a GaAs/Al_xGa_{1-x}As one-dimensional channel fabricated using an atomic force microscope. *Appl. Phys. Lett.* **78**, 3466–3468 (2001).
5. van Wees, B. J. *et al.* Quantised conductance of point contacts in a two dimensional electron gas. *Phys. Rev. Lett.* **60**, 848–850 (1988).
6. Wharam, D. A. *et al.* One-dimensional transport and the quantisation of ballistic resistance. *J. Phys. C* **21**, L209–L214 (1988).
7. Eriksson, M. A. *et al.* Cryogenic scanning probe characterization of semiconductor nanostructures. *Appl. Phys. Lett.* **69**, 671–673 (1996).
8. Woodside, M. T. & McEuen, P. L. Scanned probe imaging of single-electron charge states in nanotube quantum dots. *Science* **296**, 1098–1101 (2002).
9. Crook, R. *et al.* Quantum-dot electron occupancy controlled by a charged scanning probe. *Phys. Rev. B* **66**, 121301 (2002).
10. Ihn, T. *et al.* Scanning gate measurements on a quantum wire. *Physica E* **12**, 691–694 (2002).
11. Topinka, M. A. *et al.* Coherent branched flow in a two-dimensional electron gas. *Nature* **410**, 183–186 (2001).
12. Tessmer, S. H., Glicofridis, P. I., Ashoori, R. C., Levitov, L. S. & Melloch, M. R. Subsurface charge accumulation imaging of a quantum Hall liquid. *Nature* **392**, 51–54 (1998).
13. Tortorese, M., Barret, R. C. & Quate, C. F. Atomic resolution with an atomic force microscope using piezoresistive detection. *Appl. Phys. Lett.* **62**, 834–836 (1993).
14. Crook, R., Smith, C. G., Simmons, M. Y. & Ritchie, D. A. Imaging electrostatic microconstrictions in long 1D wires. *Physica E* **12**, 695–698 (2002).
15. Starikov, A. A. *et al.* Effects of accidental microconstriction on the quantized conductance in long wires. Preprint at (arxiv.org/abs/cond-mat/0206013) (2002).

16. Crook, R., Smith, C. G., Simmons, M. Y. & Ritchie, D. A. One-dimensional probability density observed using scanned gate microscopy. *J. Phys. Condens. Matter* **12**, L735–L740 (2000).
17. Williamson, J. G., Timmering, C. E., Harmans, C. J. P. M., Harris, J. J. & Foxon, C. T. Quantum point contact as a local probe of the electrostatic potential contours. *Phys. Rev. B* **42**, R7675–R7678 (1990).
18. Yacoby, A. *et al.* Nonuniversal conductance quantization in quantum wires. *Phys. Rev. Lett.* **77**, 4612–4615 (1996).
19. Thomas, K. J. *et al.* Interaction effects in a one-dimensional constriction. *Phys. Rev. B* **58**, 4846–4852 (1998).
20. Buttiker, M. Quantized transmission of a saddle-point constriction. *Phys. Rev. B* **41**, 7906–7909 (1990).
21. Field, M. *et al.* Measurements of Coulomb blockade with a noninvasive voltage probe. *Phys. Rev. Lett.* **70**, 1311–1314 (1993).
22. Johnson, K. L. *Contact Mechanics* (Cambridge Univ. Press, Cambridge, UK, 1987).
23. Eriksson, M. A. *et al.* Effect of a charged scanned probe microscope tip on a subsurface electron gas. *Superlattices Microstruct.* **20**, 435–440 (1996).
24. Micolich, A. P. *et al.* Evolution of fractal patterns during a classical-quantum transition. *Phys. Rev. Lett.* **87**, 036802 (2001).
25. Bennett, C. H. & DiVincenzo, D. P. Quantum information and computation. *Nature* **404**, 247–255 (2000).

Acknowledgements We thank K. J. Thomas, C. J. B. Ford and M. Kataoka for discussions. A.C.G. fabricated the sample and made the antidot measurements. C.G.S. coordinated the scanning-probe measurement facilities. I.F. and H.E.B. grew the wafer using molecular beam epitaxy (MBE). D.A.R. coordinated the MBE facilities. This work was supported by the EPSRC.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.G.S. (cgs4@cam.ac.uk).

Climate-driven changes to the atmospheric CO₂ sink in the subtropical North Pacific Ocean

John E. Dore, Roger Lukas, Daniel W. Sadler & David M. Karl

Department of Oceanography, School of Ocean and Earth Science and Technology, University of Hawaii, 1000 Pope Road, Honolulu, Hawaii 96822, USA

The oceans represent a significant sink for atmospheric carbon dioxide¹. Variability in the strength of this sink occurs on interannual timescales, as a result of regional and basin-scale changes in the physical and biological parameters that control the flux of this greenhouse gas into and out of the surface mixed layer^{2,3}. Here we analyse a 13-year time series of oceanic carbon dioxide measurements from station ALOHA in the subtropical North Pacific Ocean near Hawaii⁴, and find a significant decrease in the strength of the carbon dioxide sink over the period 1989–2001. We show that much of this reduction in sink strength can be attributed to an increase in the partial pressure of surface ocean carbon dioxide caused by excess evaporation and the accompanying concentration of solutes in the water mass. Our results suggest that carbon dioxide uptake by ocean waters can be strongly influenced by changes in regional precipitation and evaporation patterns brought on by climate variability.

The thermodynamic driving potential for the sea-to-air flux of CO₂ is the difference between the partial pressure of the dissolved CO₂ in the surface layer and that of the CO₂ in the overlying air ($\Delta p_{\text{CO}_2} = p_{\text{CO}_2\text{oc}} - p_{\text{CO}_2\text{atm}}$)⁵. The direction of the flux (F) is governed by the sign of Δp_{CO_2} , while the magnitude is also dependent on the gas exchange coefficient K , which is a parameterized function of wind speed ($F = K\Delta p_{\text{CO}_2}$)⁵. Commonly, measurements of dissolved inorganic carbon (DIC) and total alkalinity (TA) in the ocean surface mixed layer are used to calculate $p_{\text{CO}_2\text{oc}}$, while $p_{\text{CO}_2\text{atm}}$ may be obtained from nearby atmospheric sampling stations and corrected for water vapour saturation at *in situ* sea

surface temperature⁶. Beginning in October 1988, the Hawaii Ocean Time-series (HOT) project has collected, on a nearly monthly basis, a suite of physical and biogeochemical data, including surface DIC, TA, temperature, salinity and nutrient measurements, at station ALOHA (22°45' N, 158° W) in the North Pacific subtropical gyre⁴. We use these data, along with monthly $p_{\text{CO}_2\text{atm}}$ measurements from the Mauna Loa Observatory (19°32.4' N, 155°34.8' W; elevation 3,397 m) and hourly meteorological measurements from the nearby NOAA weather buoy no. 51001 (23°24.1' N, 162°11.6' W), to estimate Δp_{CO_2} , K and F from January 1989 to December 2001.

Surface ocean salinity (S), DIC and TA exhibited increasing trends with considerable interannual variability during the study period, while temperature (T) displayed an insignificant ($P > 0.05$) decreasing trend superimposed on a seasonal cycle (Fig. 1). From 1989 to 1992, S averaged 34.94‰. Substantial freshening events occurred in 1995 and 1996, driving S as low as 34.4‰. From mid-1996 to mid-1998, S rose to nearly 35.5‰, and has remained high at around 35.2–35.3‰ through to 2001. The concentrating/diluting effect of salinity changes on DIC and TA can be removed from the

data through normalization to a constant S . Following convention, values are normalized to $S = 35$ ‰ ($n\text{DIC} = 35\text{DIC}/S$ and $n\text{TA} = 35\text{TA}/S$). A strong influence of salinity on the observed trends in DIC and TA is revealed (Fig. 1b, c). The linear increase in $n\text{DIC}$ (mean $\pm$ s.e.) is $1.19 \pm 0.14 \mu\text{mol kg}^{-1} \text{yr}^{-1}$, 58% less than the trend in DIC of $2.84 \pm 0.28 \mu\text{mol kg}^{-1} \text{yr}^{-1}$; salinity normalization changes both the magnitude and sign of the TA trend, reducing it from 1.62 ± 0.29 to $-0.28 \pm 0.12 \mu\text{eq kg}^{-1} \text{yr}^{-1}$.

It is evident from Fig. 1 that while S and $n\text{TA}$ display no seasonal patterns, T and $n\text{DIC}$ have distinct annual cycles. The cycle of $n\text{DIC}$ is sinusoidal, reaching a maximum in April and a minimum in September–October (Fig. 2a). This pattern in surface waters is typical of the North Pacific subtropical gyre, and is largely due to input of DIC from below (via winter mixing) and biological draw-down of CO_2 (via photosynthesis within shallow summer mixed layers)⁷. When we calculate $p_{\text{CO}_2\text{oce}}$ from DIC and TA data, we also see a seasonal cycle, nearly opposite to that of the $n\text{DIC}$ (Fig. 2b). Increasing water temperature raises the partial pressure of dissolved gases; the observed annual cycle of $p_{\text{CO}_2\text{oce}}$ is largely due to warming in the summer/autumn and cooling in the winter/spring⁷. When $p_{\text{CO}_2\text{atm}}$ is subtracted from $p_{\text{CO}_2\text{oce}}$, the resulting Δp_{CO_2} is negative

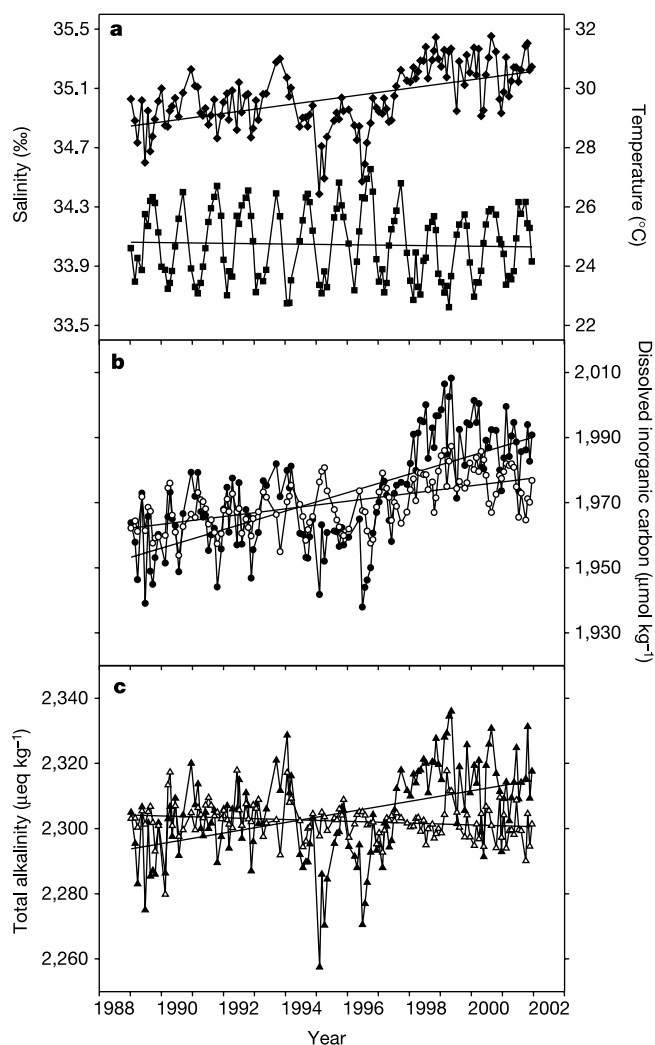


Figure 1 Variability and trends in surface mixed layer salinity (S), temperature (T), dissolved inorganic carbon (DIC) and total alkalinity (TA) at station ALOHA. **a**, Cruise mean values and linear trends in S (diamonds) and T (squares). **b**, Cruise mean values and linear trend in DIC. **c**, Cruise mean values and linear trend in TA. Filled symbols and bold lines indicate *in situ* measurements and trends; open symbols and narrow lines indicate data normalized to a constant salinity of 35‰ and corresponding trends.

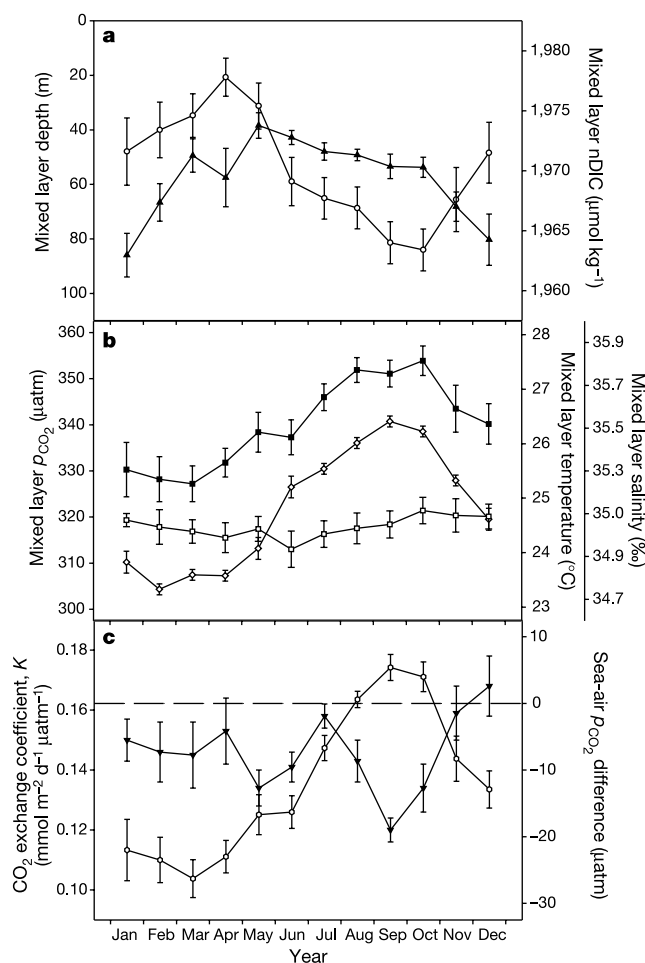


Figure 2 Seasonal patterns in important CO_2 system parameters at station ALOHA, 1989–2001. Symbols and error bars indicate mean values $\pm$ one standard error. **a**, DIC normalized to 35‰ salinity (open symbols) and depth of surface mixed layer (filled symbols; determined by 0.125 kg m^{-3} difference in potential density from the surface value). **b**, Mixed layer temperature (open diamonds) and salinity (open squares), and $p_{\text{CO}_2\text{oce}}$ (filled squares). **c**, Sea-air Δp_{CO_2} (open symbols; horizontal dashed line indicates Δp_{CO_2} of zero) and CO_2 exchange coefficient (filled symbols).

throughout most of the year, indicating a flux of CO₂ into the ocean (Fig. 2c). Only during late summer and early autumn does the flux have the potential to reverse direction; at precisely this time of year, the CO₂ gas exchange coefficient K reaches a minimum owing to low winds (Fig. 2c), thus the annual net CO₂ flux at station ALOHA has been from air to sea during each year of observation.

The concentration of atmospheric CO₂ has been rising for decades owing to anthropogenic CO₂ production, principally from the burning of fossil fuels⁸; this increase amounted to a $p_{\text{CO}_2\text{atm}}$ trend (mean $\pm$ s.e.) of $1.48 \pm 0.05 \mu\text{atm yr}^{-1}$ over the period 1989–2001 (Fig. 3a). There has been a corresponding rise in $p_{\text{CO}_2\text{oce}}$, yet of a significantly ($P < 0.05$) greater magnitude ($2.46 \pm 0.28 \mu\text{atm yr}^{-1}$). Because F depends directly upon Δp_{CO_2} , the difference between the oceanic and atmospheric p_{CO_2} trends results in a significant ($P < 0.001$) reduction with time ($0.15 \pm 0.04 \text{ mmol m}^{-2} \text{ d}^{-1} \text{ yr}^{-1}$) in the strength of the CO₂ sink (Fig. 3b). Should this reduction continue at this rate, the region would cease to act as a net sink for atmospheric CO₂ by 2008.

We may assess the effect of increasing surface salinity on this suppression of the CO₂ flux as follows. As water is evaporated, solutes are concentrated, hence both DIC and TA rise in direct proportion to salinity. The increasing DIC positively affects $p_{\text{CO}_2\text{oce}}$, while the increasing TA has a smaller negative effect. However, determination of the actual increase in $p_{\text{CO}_2\text{oce}}$ is complicated by the effects of the salinity change on the solubility of CO₂ and on the apparent dissociation coefficients for the carbonic acid system in sea water. Seawater p_{CO_2} rises by about 4.2–4.3% for a 1‰ rise in S over the range of salinities observed at station ALOHA⁹. Thus a rise in

salinity from 34.5 to 35.5‰, such as was observed from 1996 to 1998 (Fig. 1a), increases $p_{\text{CO}_2\text{oce}}$ by about 14–15 μatm . This value is roughly the same as the amplitude of the annual cycle of Δp_{CO_2} (Fig. 2), hence the salinity effect substantially influences the strength of the sea–air CO₂ flux in this region.

Through integration of F over time, we determine that the net annual sea–air flux of CO₂ at station ALOHA between 1989 and 2001 varied from -1.02 to -0.17 mol m^{-2} (Fig. 4a). When we recalculate the $p_{\text{CO}_2\text{oce}}$ data from Fig. 3 while holding S constant at 34.94‰ (the initial 1989–92 mean), it becomes evident that the CO₂ sink was strongly affected by salinity, particularly from 1996 to 2001 (Fig. 4a). Annual CO₂ fluxes were suppressed by as much as 0.29 mol m^{-2} relative to a constant-salinity scenario (Fig. 4b). If the effects of changing salinity are removed from the data, 40% of the observed 13-yr trend in F disappears. The remaining trend of $0.09 \pm 0.04 \text{ mmol m}^{-2} \text{ d}^{-1} \text{ yr}^{-1}$ is still significant ($P < 0.05$), and

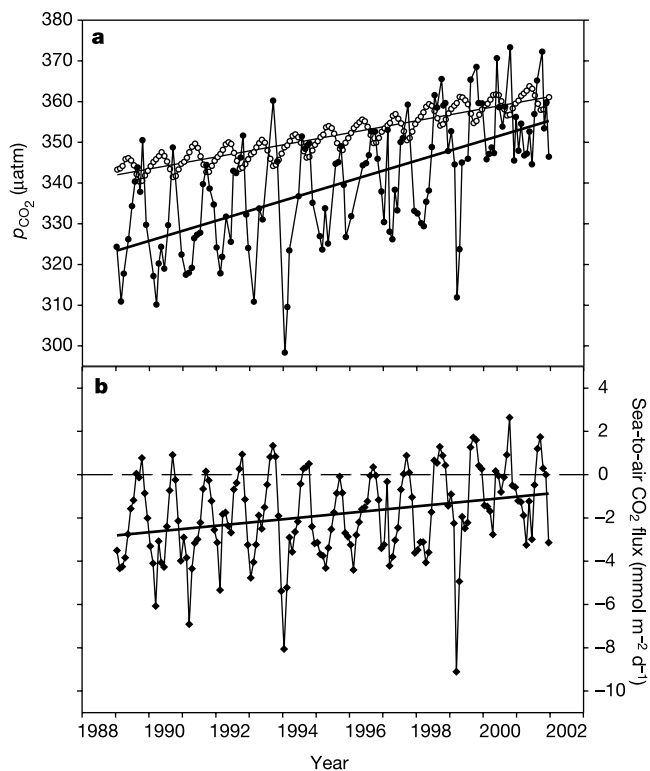


Figure 3 Interannual variability and trends in $p_{\text{CO}_2\text{atm}}$, $p_{\text{CO}_2\text{oce}}$ and the sea–air flux of CO₂ at station ALOHA. **a**, Cruise mean values and linear trend in p_{CO_2} . Filled symbols and solid line indicate ocean mixed layer p_{CO_2} measurements and trend; open symbols and narrow line indicate atmospheric data corrected for water saturation and corresponding trend. **b**, Monthly flux of CO₂ from ocean to atmosphere. Values of $p_{\text{CO}_2\text{oce}}$ were interpolated to the fifteenth day of each month to facilitate calculation of monthly F . Solid line indicates linear trend in flux.

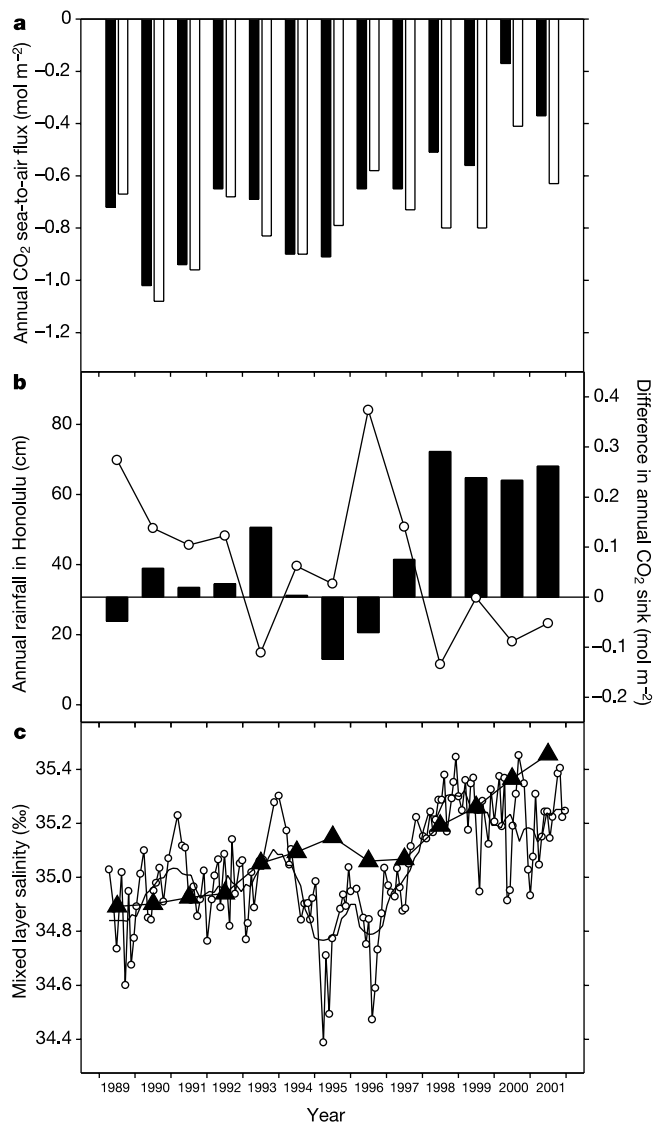


Figure 4 Salinity effect on annual net sea–air fluxes of CO₂ at station ALOHA. **a**, Measured fluxes (filled bars) and fluxes predicted at a constant salinity of 34.94‰ (open bars). **b**, Reduction in CO₂ flux attributable to salinity effect (filled bars) and total annual rainfall at Honolulu International Airport (open symbols). **c**, Mixed layer salinity (open symbols) and its annual running mean (bold curve), and salinity predicted (solid symbols) by simple rainfall model (see text).

may be due to the effects of interannual variations in mixing, advection or biological processes. For example, the observed decreasing trend in nTA (Fig. 1c) contributes to rising $p_{\text{CO}_2\text{oce}}$, and may indicate long-term changes in the rates of biological calcification and/or calcite export and dissolution. Because T displays no rising temporal trend (Fig. 1a), temperature variations cannot be invoked to account for the remaining decrease in sink strength.

When the magnitude of the salinity effect and the annual rainfall at Honolulu International Airport (21°18' N, 157°54' W) are compared, a correspondence between low rainfall and suppression of the CO_2 sink is suggested (Fig. 4b). Our shipboard sampling gives us only snapshots in time; the measured salinity on any particular cruise will be affected strongly by the recent history of rainfall and mixed layer depth. However, we can roughly estimate the effect of rainfall deficits on surface salinity as follows. First, we assume that the deepest mixed layers observed (~120 m) occur at some time during each winter, and that the annual rainfall at Honolulu is indicative of the rainfall mixed to that depth at station ALOHA. Second, we assume that evaporation and precipitation were roughly in balance over the initial 1989–92 time period, and that evaporation remained constant thereafter (53.53 cm yr^{-1}). Although the rainfall at station ALOHA may not always track that in Honolulu, and although variations in evaporation, turbulent salt flux and water mass advection must be considered for a rigorous salt budget, the results of this exercise suggest that a net freshwater loss from the ocean due to the observed rainfall deficit is sufficient to account for the magnitude of the observed salinity increase (Fig. 4c).

We have shown that rising salinity in the subtropical North Pacific Ocean near Hawaii from 1989 to 2001, substantially influenced by recent rainfall deficits, has directly caused a reduction in the strength of the CO_2 sink. Most of this salinity increase occurred over a two-year period (1996–98), hence we may be observing a manifestation of a climate oscillation or regime shift, rather than a long-term temporal trend. Drought conditions over a large area of the North Pacific have persisted since the mid-1990s^{10–12}, and may be related to both the El Niño and Pacific Decadal Oscillation climate phenomena^{11–13}. In addition, recent biogeochemical and fisheries data suggest a logistical response to a potential late-1990s shift in North Pacific climate^{14,15}. As the ocean CO_2 time-series data records lengthen, our ability to draw direct links between specific climate phenomena and interannual to decadal variability in the oceanic carbon cycle will improve. Regional changes in evaporation and precipitation patterns accompanying climate variability have the potential to significantly alter the global ocean sink for CO_2 , and should be considered when modelling the ocean carbon cycle. Moreover, there is reason to be concerned about the influence of hydrologic changes on the response of the oceanic CO_2 sink to anthropogenic climate change. □

Methods

Sampling and analyses

Shipboard sampling was conducted within a six-nautical-mile radius of the nominal position of station ALOHA. Discrete water samples were collected using a rosette system fitted with a SeaBird CTD (conductivity, temperature, depth) sensor for continuous salinity and temperature determinations. Samples for DIC and TA were returned to our shore-based laboratory, and analysed using semi-automated coulometry and open-cell titration techniques, respectively, following the recommendations of the DOE¹⁶ with minor modifications. The accuracy of these measurements was maintained using certified seawater reference materials¹⁷. Calculations of $p_{\text{CO}_2\text{oce}}$ were carried out¹⁸, using independently validated^{19,20} parameterizations of the apparent dissociation constants for carbonic acid in sea water²¹ that were subsequently refitted²².

Supporting data sets

Monthly atmospheric CO_2 mole fractions measured at the Mauna Loa Observatory²³ were obtained from the Carbon Dioxide Information Analysis Center (<http://cdiac.ornl.gov/ftp/maunaloa-co2/maunaloa.co2>). Monthly rainfall data from Honolulu International Airport were obtained from the National Weather Service Pacific Regional Headquarters (http://www.prh.noaa.gov/hnl/climate/PHNL_rainfall.html). Hourly measurements of sea surface temperature and wind speed from weather buoy no. 51001 were obtained from

the National Weather Service National Data Buoy Center (http://seaboard.ndbc.noaa.gov/station_history.phtml?station=51001). These meteorological data sets have been shown to be well correlated with discrete observations made at station ALOHA during HOT cruises²⁴. Hourly buoy data was used to calculate K (ref. 5), which was then binned by month.

Treatment of missing data

The HOT programme obtained measurements of mixed layer DIC and TA on 122 and 118 of the 131 cruises completed during 1989–2001, respectively. Several additional data points during 1989–90 were obtained from parallel measurements²⁵ made on samples collected during HOT cruises using similar methods and identical reference materials. On a single cruise (HOT-4; February 1989), a DIC measurement was available, but an average value of $2,303.2 \mu\text{eq kg}^{-1}$ was assumed for nTA. In all, $p_{\text{CO}_2\text{oce}}$ was calculated for 125 occupations of station ALOHA between 1989 and 2001. Several short gaps in the buoy data sets were filled using the monthly averaged values from the entire 13-yr record.

Received 17 February; accepted 3 July 2003; doi:10.1038/nature01885.

1. Siegenthaler, U. & Sarmiento, J. L. Atmospheric carbon dioxide and the ocean. *Nature* **365**, 119–125 (1993).
2. Bates, N. R., Pequignat, A. C., Johnson, R. J. & Gruber, N. A short-term sink for atmospheric CO_2 in subtropical mode water of the North Atlantic Ocean. *Nature* **420**, 489–493 (2002).
3. Gruber, N., Keeling, C. D. & Bates, N. R. Interannual variability in the North Atlantic Ocean carbon sink. *Science* **298**, 2374–2378 (2002).
4. Karl, D. M. & Lukas, R. The Hawaii Ocean Time-series (HOT) program: Background, rationale and field implementation. *Deep-Sea Res. II* **43**, 129–156 (1996).
5. Wanninkhof, R. Relationship between wind speed and gas exchange over the ocean. *J. Geophys. Res.* **97**, 7373–7382 (1992).
6. Winn, C. D., Mackenzie, F. T., Carrillo, C. J., Sabine, C. L. & Karl, D. M. Air-sea carbon dioxide exchange in the North Pacific Subtropical Gyre: Implications for the global carbon budget. *Glob. Biogeochem. Cycles* **8**, 157–163 (1994).
7. Ishii, M. *et al.* Seasonal variation in total inorganic carbon and its controlling processes in surface waters of the western North Pacific subtropical gyre. *Mar. Chem.* **75**, 17–32 (2001).
8. Houghton, J. T., Jenkins, G. J. & Ephraums, J. J. (eds) *Climate Change: The IPCC Scientific Assessment* (Cambridge Univ. Press, Cambridge, UK, 1990).
9. Weiss, R. F., Jahnke, R. A. & Keeling, C. D. Seasonal effects of temperature and salinity on the partial pressure of CO_2 in seawater. *Nature* **300**, 511–513 (1982).
10. Waple, A. M. *et al.* Climate assessment for 2001. *Bull. Am. Meteorol. Soc.* **83**, S1–S62 (2002).
11. Minobe, S. & Nakanowatari, T. Global structure of biennial precipitation variability in boreal winter. *Geophys. Res. Lett.* **29**, 1396–1399 (2002).
12. Lukas, R. Freshening of the upper thermocline in the North Pacific subtropical gyre associated with decadal changes in rainfall. *Geophys. Res. Lett.* **28**, 3485–3488 (2001).
13. Hoerling, M. & Kumar, A. The perfect ocean for drought. *Science* **299**, 691–694 (2003).
14. Dore, J. E., Brum, J. R., Tupas, L. M. & Karl, D. M. Seasonal and interannual variability in sources of nitrogen supporting export in the oligotrophic subtropical North Pacific Ocean. *Limnol. Oceanogr.* **47**, 1595–1607 (2002).
15. Chavez, F. P., Ryan, J., Lluch-Cota, S. E. & Niquen, M. C. From anchovies to sardines and back: Multidecadal change in the Pacific Ocean. *Science* **299**, 217–221 (2003).
16. DOE *Handbook of Methods for the Analysis of the Various Parameters of the Carbon Dioxide System in Sea Water* Version 2 (eds Dickson, A. G. & Goyet, C.) ORNL/CDIAC-74 (Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, US Department of Energy, Oak Ridge, Tennessee, 1994).
17. Dickson, A. G. Reference materials for oceanic CO_2 measurements. *Oceanography* **14**, 21–22 (2001).
18. Lewis, E. & Wallace, D. W. R. *Program Developed for CO_2 System Calculations* ORNL/CDIAC-105 (Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, US Department of Energy, Oak Ridge, Tennessee, 1998).
19. Lueker, T. J., Dickson, A. G. & Keeling, C. D. Ocean p_{CO_2} calculated from dissolved inorganic carbon, alkalinity, and equations for K_1 and K_2 : Validation based on laboratory measurements of CO_2 in gas and seawater at equilibrium. *Mar. Chem.* **70**, 105–119 (2000).
20. Lee, K., Millero, F. J., Byrne, R. H., Feely, R. A. & Wanninkhof, R. The recommended dissociation constants for carbonic acid in seawater. *Geophys. Res. Lett.* **27**, 229–232 (2000).
21. Merbach, C., Culbertson, C. H., Hawley, J. E. & Pytkowicz, R. M. Measurement of the apparent dissociation constants of carbonic acid in seawater at atmospheric pressure. *Limnol. Oceanogr.* **18**, 897–907 (1973).
22. Dickson, A. G. & Millero, F. J. A comparison of the equilibrium constants for the dissociation of carbonic acid in seawater media. *Deep-Sea Res.* **34**, 1733–1743 (1987).
23. Keeling, C. D. & Whorf, T. P. in *Trends: A Compendium of Data on Global Change* (Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, US Department of Energy, Oak Ridge, Tennessee, 2002).
24. Tupas, L., *et al.* *Hawaii Ocean Time-series Data Report 4: 1992* (SOEST Technical Report 93-14, Univ. Hawaii, 1993).
25. Keeling, C. D. in *The Global Carbon Cycle* (ed. Heimann, M.) 413–429 (NATO ASI Series, Vol. I15, Springer, Berlin, 1993).

Acknowledgements We thank past and present HOT programme scientists and staff, seagoing support personnel and the crews of the many research vessels used to collect these data. We also thank the National Oceanic and Atmospheric Administration and the Department of Energy for access to critical meteorological and atmospheric CO_2 data sets. This work was supported by the National Science Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.E.D. (jdore@hawaii.edu).

Stabilizing feedbacks in glacier-bed erosion

R. B. Alley¹, D. E. Lawson², G. J. Larson³, E. B. Evenson⁴ & G. S. Baker⁵

¹Department of Geosciences and EMS Environment Institute, Pennsylvania State University, University Park, Pennsylvania 16802, USA

²Cold Regions Research and Engineering Laboratory, Hanover, New Hampshire 03755, USA

³Department of Geological Sciences, Michigan State University, East Lansing, Michigan 48824, USA

⁴Department of Earth and Environmental Sciences, Lehigh University, Bethlehem, Pennsylvania 18015, USA

⁵Department of Geology, University at Buffalo, Buffalo, New York 14260, USA

Glaciers often erode, transport and deposit sediment much more rapidly than nonglacial environments¹, with implications for the evolution of glaciated mountain belts and their associated sedimentary basins. But modelling such glacial processes is difficult, partly because stabilizing feedbacks similar to those operating in rivers^{2,3} have not been identified for glacial landscapes. Here we combine new and existing data of glacier morphology and the processes governing glacier evolution from diverse settings to reveal such stabilizing feedbacks. We find that the long profiles of beds of highly erosive glaciers tend towards steady-state angles opposed to and slightly more than 50 per cent steeper than the overlying ice–air surface slopes, and that additional subglacial deepening must be enabled by non-glacial processes. Climatic or glaciological perturbations of the ice–air surface slope can have large transient effects on glaciofluvial sediment flux and apparent glacial erosion rate.

Geomorphologists have long used the concept of a graded river, in which “...any change in any of the controlling factors will cause a displacement of the equilibrium in a direction that will tend to absorb the effect of the change”². For example, increased sediment supply or decreased water supply to a segment of a river initially in steady state causes sediment supply to exceed sediment transport there; the resulting sediment accumulation steepens the river bed in the downstream direction, increasing transport to restore balance.

Graded rivers are a consequence of the coupling of mass conservation with a slope-dependent mass-transport law. Nearly complete independence of ice flow from bed slope has prevented a similar concept for glaciers. However, a thermodynamic effect introduces dependence of subglacial transport on a combination of surface and bed slope, stabilizing beds of terminal regions of highly erosive glaciers at a slope just sufficient to cause supercooling of subglacial waters. Process understanding motivates this ‘graded glacier’ hypothesis, which is supported by data indicating widespread supercooling, and both increase and decrease of bed slopes of different glaciers towards the equilibrium condition.

Considering processes first, rapid glacial geomorphic activity appears to be restricted to glaciers with abundant surface meltwater flowing along their beds, especially if those glaciers are large and slide rapidly^{1,4,5}. The sediment discharge from such highly erosive glaciers is dominated by material carried in subglacial streams^{6–8}. In the absence of such drainage, any rapid erosion would quickly produce a till (the equivalent of a fault gouge) that would protect bedrock from the ice and greatly limit further erosion^{9–11}. (Widespread and sustained entrainment of sediment to basal ice sufficiently rapidly to balance the faster rates of glacial erosion is thermodynamically difficult and has not been observed (see review⁸)).

Rapid subglacial erosion produces overdeepenings, with the glacier bed rising in the direction of ice flow. Overdeepenings may form in cirques near glacier heads or anywhere along the length of a glacier¹², but are prominent in downglacier reaches, probably owing

to the downglacier increase in surface meltwater reaching the bed and its critical role in erosion¹³.

The work done in moving subglacial water is partitioned between heat generation from viscous dissipation, and increase in potential energy of the water during ascent from any overdeepening(s)^{14–16}. Some heat is required to warm the water as the melting point rises in response to falling pressure along water flow. For beds decreasing in elevation or increasing only gradually in the direction of ice and water flow, the potential-energy term is small or negative, heat production from viscous dissipation exceeds that needed to maintain water at the pressure-melting temperature, most of the excess heat melts the ice walls of water channels, and the associated channel expansion produces lower water pressure. Heat is generated throughout the channel cross-sectional area, which increases with channel radius more rapidly than does the perimeter being melted; thus, melting is faster and produces lower pressure in larger channels, in turn favouring downglacier convergence of water flow¹⁴. Because sediment-transport capacity increases more rapidly than water flux^{3,8}, convergent water flow favours erosion and transport, accelerating local deepening of the glacier bed and formation and steepening of the downglacier sides of overdeepenings (Fig. 1a).

However, if the bed rises at a sufficiently steep angle in the direction of flow, then enough of the work done in moving water is stored as potential energy that viscous dissipation is not sufficient to warm the water along the pressure-melting curve. The critical angle, or ‘supercooling threshold’, is ~20–70% steeper than and opposed to the ice–air surface slope, with the range arising from uncertainty about maintenance of air-saturation of the water under changing pressure^{12,14,17}. Beds steeper than the threshold cause supercooling and ice growth, releasing latent heat. Some of the ice grows on channel walls, constricting channels, raising water pressure, and causing divergent water flow into a distributed ice-contact drainage system; the resulting decrease in sediment-transport capacity along flow favours deposition. This will eliminate local erosion, trap sediment transported from upglacier (Fig. 1b), and so prevent further local deepening of the glacier bed and steepening of the slope on the downglacier side of the overdeepening^{5,18}.

Process understanding thus shows that erosion faster than in the proglacial environment by a large, meltwater-rich glacier can continue until the slope of the downglacier side of a terminal overdeepening steepens across the supercooling threshold, interfering with further erosion and sediment evacuation. Steepening of the ice–air surface slope will allow additional erosional steepening of the subjacent bed slope, and flattening of the ice–air surface slope will tend to trap sediment transported from upglacier, thus maintaining weak supercooling and ice growth near the supercooling threshold. Averaged over an appropriately long interval, which will depend on the nonsteadiness of climate and on the response time of the glacial-geomorphic system, further downcutting beneath terminal regions of a glacier near the supercooling threshold should depend on the ability of the proglacial environment to erode downward or on isostatic or tectonic tilting. Proglacial erosion rates or tectonic processes thus should control subglacial downcutting by long-lived, highly erosive glaciers. Additional glacial erosion would also remain possible through headward growth of cirques¹⁹, valley widening⁴, and downcutting of portions of the valley not yet at the supercooling threshold, including formation or upglacier extension of terminal or of upglacier overdeepenings (which have slopes on their downglacier sides that probably also tend to the supercooling threshold¹³).

Our compilation of new and existing data from diverse glacial settings to test this graded-glacier hypothesis provides much support. First, supercooling is active beneath many glaciers, including some in Sweden¹⁵, Alaska, USA¹⁸, and Iceland²⁰.

Additionally, steepening of glacier beds to stabilize near the supercooling threshold is illustrated by the recent history of the

Taku and Muir glaciers in Alaska. Both experienced post-Little Ice Age retreats along fjords that left the glaciers without large proglacial morainial shoals^{21–23}. Both then discharged very large quantities of sediment into their proglacial fjords, rapidly building prominent moraine shoals with upglacier slopes that evolved to angles just steep enough to cause supercooling^{5,21,24}, and exhibiting evidence that supercooling was active⁵. Erosion of overridden sediments of 3 m yr^{-1} beneath Taku²¹ and proglacial sedimentation of 30 m yr^{-1} at the nearby, similar Riggs glacier in Alaska²⁵ illustrate the spectacular rates at which sediment can be moved as overdeepenings form and steepen towards the supercooling threshold; even for erosion into bedrock, rates of 10 mm yr^{-1} or more may be possible¹.

Rapid reduction of the ice–air surface slope can shift a glacier far into the supercooling condition. Evidence from Matanuska glacier, Alaska, indicates that in response to such a slope reduction, sedimentation is filling overdeepenings and reducing bed slopes

back towards the supercooling threshold. The best-studied, prominent terminal overdeepening of Matanuska glacier now has a bed slope opposed to and typically several times steeper than the ice–air surface slope, producing extensive supercooling of subglacial waters¹⁸. (The ice–air surface slope is even locally reversed over the upglacier end of this overdeepening, causing lakes to form on the ice.) Average ice–air surface slope overlying the overdeepening has probably decreased over time in response to changing climate; observations of Matanuska glacier in 1954 and photographs taken in 1898 have been used²⁶ to demonstrate that considerable thinning but little horizontal retreat had occurred in downglacier reaches over the 56-yr interval (subsequent retreat has shifted the active ice margin a few hundred metres from its former moraine).

Common cobbles on an outwash surface with its head graded to the former moraine show that abundant bed load was discharged in the past, probably through subglacial channels that have been disrupted during the subsequent surface-slope reduction into the supercooling field. The metres-thick silt-bearing basal ice that has accreted to Matanuska glacier from supercooled subglacial discharge^{18,27} occasionally contains evidence of basal or near-basal water channels plugged by radial growth of ice crystals from channel walls²⁸. Dye-tracing and other data sets from the best-characterized overdeepening indicate a distributed, high-pressure, low-velocity basal water system despite water entering the glacier in vigorous streams at large moulins¹⁸.

Such a distributed water system would not be expected to carry much coarse sediment. Consistent with this, observations during one summer at two prominent vents²⁹ found almost zero discharge of coarse-grained bed load ($<1\%$ of total load versus subequal suspended and bed loads for typical glaciers⁶). Subaqueous fan deposits at these vents, observed in the autumn after melting slowed before snowfall, probably contained the coarsest material discharged during the entire melt season, yet were fine-grained, with pebble-sized or larger clasts being rare.

Suppression of erosion beneath the terminal region may contribute to this lack of bed load; however, most of the glacier is at higher elevations than the terminal overdeepenings (the glacier is about 45 km long, with $3,000 \text{ m}$ of vertical relief, and an unglaciated area of 185 km^2 in the drainage basin potentially contributing much coarse material by inwash or infall to the 380 km^2 glacier¹⁸), so it is highly likely that cobble-sized and coarser clasts are being generated and delivered to the terminal overdeepening. Furthermore, coarse clasts are present in the terminal overdeepening, as shown by limited cobbles in the silt-rich basal accretion ice and by the clast-bearing, although fine-grained, till that is frozen to the glacier sole by ice-marginal wintertime conductive cooling and exposed by thrusting of the ice margin²⁷. And, abundant cobbles were discharged to the outwash surface before the decrease in ice–air surface slope shifted the glacier so far into the supercooling field.

The total effect of the supercooling on the sediment budget is difficult to estimate, in large part because we lack pre-supercooling sediment-discharge data for Matanuska glacier, and because the subglacial sediment trapping does not appear to be fast enough for easy direct measurement. In light of the common dominance of stream discharge in sediment budgets (for example, ref. 7), typically with subequal bedload and suspended load⁶, the nearly total lack of Matanuska bedload discharge may indicate a supercooling-caused reduction in sediment discharge of the order of 50% . The suspended sediment still discharging through subglacial water is equivalent to roughly 1 mm yr^{-1} of bedrock erosion averaged over the complete glacier area³⁰. Setting the trapped bed load equal to this discharged suspended load, and assuming that terminal overdeepening(s) occupy $<10\%$ of the glacier area, we infer sedimentation in terminal overdeepening(s) of order 10 mm yr^{-1} or more, tending to return the bed slope closer to the supercooling threshold.

Because the divergent water flow from ice growth clogging subglacial streams is expected to reduce suspended-load as well as

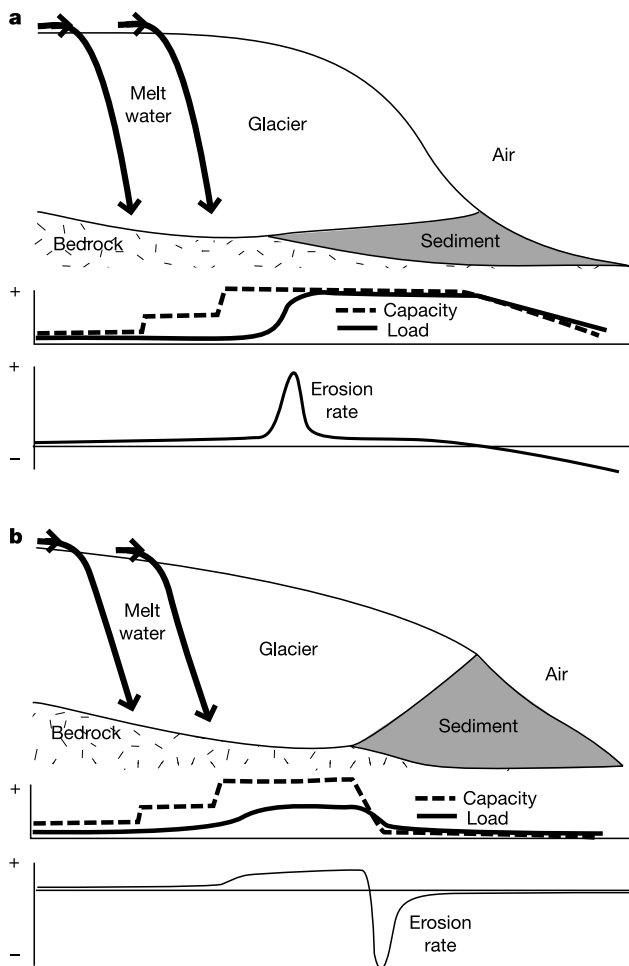


Figure 1 Cartoons of sediment-transport capacity and sediment load in subglacial streams, for a glacier lacking supercooling (**a**), and for a glacier far into the supercooling field (**b**). In each case, a moraine or moraine shoal has developed, so that the terminal overdeepening is in part a sediment-floored feature. In **a**, load is less than capacity in upglacier regions owing to the difficulty of eroding bedrock, but load rises to match capacity on the overridden and easily erodible sediments, as erosion steepens the upglacier slope of the sediments. In **b**, ice growth from supercooling of streams flowing up the overly steep face of the moraine shoal causes transport capacity to drop below load delivered, producing deposition (shown as a negative erosion rate in the lower panel of **b**) to fill the overdeepening back towards the supercooling threshold. Were the proglacial environment able to remove all the sediment delivered but not able to erode bedrock as rapidly as the glacier, then the upper panel of **b** could also be drawn with ice wholly on bedrock, and subglacial erosion lowering the glacier bed below the proglacial environment.

bedload transport capacity⁸, Matanuska glacier might be trapping fine-grained as well as coarse-grained sediment subglacially. This is consistent with the observed fine-grained character of the subglacial tills, and would mean that subglacial sedimentation is faster than estimated above. However, estimates are complicated by sediment trapped in the ice accreted to the base of the glacier and then discharged by ice-marginal melting (which may transport 0.1–1 mm yr⁻¹ of erosion averaged over the glacier area), as well as by any sediment transport through subglacial till deformation. Certainly, given the very high sediment fluxes from some glaciers, and the possibility that supercooling can trap a substantial fraction of the sediment flux in spatially restricted overdeepenings, geomorphically and glaciologically important sedimentation rates in those basins seem likely.

We thus hypothesize that beds of highly erosive glaciers tend to evolve to an equilibrium angle close to but slightly steeper than the supercooling threshold, which is 20–70% steeper than and opposed to the ice–air surface slope. Additional data are clearly desirable, but this hypothesis is supported by our understanding of processes, by the widespread occurrence of overdeepenings with supercooling, and by evidence indicating evolution of both too-gradual and too-steep bed slopes towards the supercooling threshold. Some supercooling is likely at equilibrium to suppress the tendency for continued excavation of overdeepenings; the exact equilibrium slope should depend on several factors including the supply to terminal overdeepenings of sediment that must be moved in steady state, whether the sediment is from nonglacial sources, from valley widening or headward extension, or from downcutting in regions not yet at the supercooling threshold.

Our hypothesis implies that, over sufficiently long times, the bed profiles of terminal regions of highly erosive glaciers will be controlled by the ability of proglacial regions to downcut, or by tectonic or isostatic tilting. Over shorter times, perturbations to the surface slope can greatly affect sediment transport and erosion (by twofold to perhaps an order of magnitude or more), with interesting corollaries such as the possibility that kinematic waves in the ice could serve as sediment pumps. Proglacial sediments should thus be sensitive recorders of climatic and glaciological changes, but interpreting that record will require careful consideration of the glacial sedimentary system. □

Received 16 February; accepted 23 June 2003; doi:10.1038/nature01839.

- Hallet, B., Hunter, L. & Bogen, J. Rates of erosion and sediment evacuation by glaciers: a review of field data and their implications. *Glob. Planet. Change* **12**, 213–235 (1996).
- Mackin, J. H. Concept of the graded river. *Geol. Soc. Am. Bull.* **59**, 463–512 (1948).
- Easterbrook, D. J. *Surface Processes and Landforms* 2nd edn (Prentice Hall, Upper Saddle River, NJ, 1999).
- Montgomery, D. R. Valley formation by fluvial and glacial erosion. *Geology* **30**, 1047–1050 (2002).
- Alley, R. B., Lawson, D. E., Evenson, E. B. & Larson, G. J. Sediment, glaciohydraulic supercooling, and fast glacier flow. *Ann. Glaciol.* (in the press).
- Lawson, D. E. *Glaciohydraulic and Glaciological Effects on Runoff and Sediment Yield in Glacierized Basins* (CRREL Monograph 93–2, US Army Cold Regions Research and Engineering Laboratory, Hanover, NH, 1993).
- Hunter, L. E., Powell, R. D. & Lawson, D. Flux of debris transported by ice at three Alaskan tidewater glaciers. *J. Glaciol.* **42**, 123–135 (1996).
- Alley, R. B. *et al.* How glaciers entrain and transport sediment at their beds: Physical constraints. *Quat. Sci. Rev.* **16**, 1017–1038 (1997).
- Iverson, N. R. Potential effects of subglacial water-pressure fluctuations on quarrying. *J. Glaciol.* **37**, 27–36 (1991).
- Hallet, B. Glacial quarrying: a simple theoretical model. *Ann. Glaciol.* **22**, 1–9 (1996).
- Cuffey, K. M. & Alley, R. B. Erosion by deforming subglacial sediments: Is it significant? (Toward till continuity). *Ann. Glaciol.* **22**, 126–133 (1996).
- Hooke, R. & Le, B. Positive feedbacks associated with erosion of glacial cirques and overdeepenings. *Geol. Soc. Am. Bull.* **34**, 228–231 (1991).
- Alley, R. B., Strasser, J. C., Lawson, D. E., Evenson, E. B. & Larson, G. J. in *Glacial Processes Past and Present* (eds Mickelson, D. M. & Attig, J. W.) 1–9 (Geol. Soc. Am. Spec. Pap. 337, Boulder, CO, 1999).
- Röthlisberger, H. & Lang, H. in *Glaciofluvial Sediment Transfer—An Alpine Perspective* (eds Gurnell, A. M. & Clark, M. J.) 207–284 (Wiley, New York, 1987).
- Hooke, R., Le, B. & Pohjola, V. A. Hydrology of a segment of a glacier situated in an overdeepening, Storglaciären, Sweden. *J. Glaciol.* **40**, 140–148 (1994).
- Alley, R. B., Lawson, D. E., Evenson, E. B., Strasser, J. C. & Larson, G. J. Glaciohydraulic supercooling: A freeze-on mechanism to create stratified, debris-rich basal ice. 2. Theory. *J. Glaciol.* **44**, 563–569 (1998).
- Lliboutry, L. Modifications to the theory of intraglacial waterways for the case of subglacial ones. *J. Glaciol.* **29**, 216–226 (1983).

- Lawson, D. E. *et al.* Glaciohydraulic supercooling: A freeze-on mechanism to create stratified, debris-rich basal ice. 1. Field evidence and conceptual model. *J. Glaciol.* **44**, 547–562 (1998).
- Brocklehurst, S. H. & Whipple, K. X. Glacial erosion and relief production in the Eastern Sierra Nevada, California. *Geomorphology* **42**, 1–24 (2002).
- Roberts, M. J. *et al.* Glaciohydraulic supercooling in Iceland. *Geology* **30**, 439–442 (2002).
- Nolan, M., Motyka, R. J., Echelmeyer, K. & Trabant, D. C. Ice-thickness measurements of Taku Glacier, Alaska, USA, and their relevance to its recent behavior. *J. Glaciol.* **41**, 541–553 (1995).
- Hall, D. K., Benson, C. S. & Field, W. O. Changes of glaciers in Glacier Bay, Alaska, using ground and satellite measurements. *Phys. Geogr.* **16**, 27–41 (1995).
- Koppes, M. N. & Hallet, B. Influence of rapid glacial retreat on the rate of erosion by tidewater glaciers. *Geology* **30**, 47–50 (2002).
- Arcone, S. A., Lawson, D. E., Delaney, A. J. & Moran, M. in *8th Int. Conf. Ground-Penetrating Radar* (eds Noon, D. A., Stickley, G. F. & Longstaff, D.) 377–382 (SPIE, Bellingham, WA, 2000).
- Powell, R. D. in *Glacimarine Environments, Processes and Sediments* (eds Dowdeswell, J. A. & Scourse, J. D.) 53–73 (Geol. Soc. Lond. Spec. Pub. No. 53, London, 1990).
- Williams, J. R. & Ferrians, O. J. Late Wisconsinan and recent history of the Matanuska Glacier, Alaska. *Arctic* **14**, 82–90 (1961).
- Lawson, D. E. *Sedimentological Analysis of the Western Terminus Region of the Matanuska Glacier*, Alaska (CRREL Report 79–9, US Army Cold Regions Research and Engineering Laboratory, Hanover, NH, 1979).
- Evenson, E. B. *et al.* in *Glacial Processes Past and Present* (eds Mickelson, D. M. & Attig, J. W.) 23–25 (Geol. Soc. Am. Spec. Pap. 337, Boulder, CO, 1999).
- Pearce, J. T. *et al.* Bedload component of glacially discharged sediment: Insights from the Matanuska Glacier, Alaska. *Geology* **31**, 7–10 (2003).
- Linker, J. S. *et al.* Suspended sediment flux as an indicator for rate of glacial erosion at the Matanuska Glacier, south-central Alaska. *Geol. Soc. Am. Abstr. Prog.* **32**, A329 (2000).

Acknowledgements We thank K. Cuffey for advice and for comments on this manuscript, including help with the third paragraph. We thank our many colleagues in studying the Matanuska glacier system, including S. Arcone, S. Kocczynski, R. Bigl, J. Denner, S. Ensminger, T. Johnston and J. Strasser, and we thank R. Hooke for suggestions. Discussions with B. Hallet, G. Clarke, T. Creyts, S. Tulaczyk and other colleagues helped clarify our ideas. We thank NSF and CRREL for funding.

Author contributions R.B.A. led collaborative theory development and writing, D.E.L., G.J.L. and E.B.E. provided extensive data and insights on glacier processes, and G.S.B. produced new geophysical data at Matanuska glacier.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.B.A. (ralley@essc.psu.edu).

Intense equatorial flux spots on the surface of the Earth's core

Andrew Jackson

School of Earth Sciences, University of Leeds, Leeds LS2 9JT, UK

A large number of high-accuracy vector measurements of the Earth's magnetic field have recently become available from the satellite Oersted, complementing previous vector data from the satellite Magsat, which operated in 1979/80. These data can be used to infer the morphology of the magnetic field at the surface of the fluid core¹, ~2,900 km below the Earth's surface. Here I apply a new methodology to these data to calculate maps of the magnetic field at the core surface which show intense flux spots in equatorial regions. The intensity of these features is unusually large—some have intensities comparable to high-latitude flux patches near the poles, previously identified as the major component of the dynamo field². The tendency for pairing of some of these spots to the north and south of the geographical equator suggests they might be associated with the tops of equatorially symmetric columnar structures in the fluid, or their antisymmetric equivalents. The drift of the equatorial features may represent material flow or could represent wave motion; discrimination of these two effects based on future data could provide new information on the strength of the hidden toroidal magnetic field of the Earth.

The Earth's magnetic field is generated in the fluid core of the Earth (radius $a = 3,485$ km) by a 'self-exciting dynamo' mechanism. This mechanism has sustained the field from decay over most of the lifetime of the Earth, and convective motions associated with the dynamo mechanism are responsible for the 'secular variation', or the slow changes in the magnetic field over timescales of decades to centuries which are seen on the Earth's surface.

Smooth images of the core field can be reliably constructed back to the 16th century using data collected on land and sea³. As in virtually any inverse problem, the problem of reconstructing the core magnetic field from surface observation is formally non-unique^{4,5}, and therefore requires some prior information in its solution; generally a quadratic smoothness criterion is implemented on the magnetic field at the core surface, which therefore roughly measures magnetic energy. Such a methodology is termed regularization and is common to most inverse problems^{5,6}. There is no reason to suppose that any quadratic property of the field should remain constant. However, both theory and numerical simulations of the geodynamo support the view that on the decade-to-century timescale the approximation of the core as a perfect electrical conductor is a good one. Thus, Alfvén's celebrated frozen-flux theorem will be obeyed to a high degree, and indeed numerical estimates put the level of violation at a few per cent per century⁷, whereas observationally there is no evidence that requires violation. When faced with making images of the core field at different points in time from sometimes vastly different data sets, an appealing method would be one based on frozen-flux: such images would

visually have similar complexity in time, while conserving a physically relevant quantity.

Bondi and Gold⁸ first drew attention to a consequence of Alfvén's theorem, namely that the 'unsigned flux N ' over the entire core surface Ω :

$$N = \int_{\Omega} |B_r| d\Omega$$

must be invariant with time. The field must also obey the condition:

$$\int_{\Omega} B_r d\Omega = 0$$

since there are no monopoles allowed by Maxwell's equations. These two equations imply that the amount of positive flux (the 'red' flux in Fig. 1), which we call B_r^+ , and the amount of negative flux (the 'blue' flux), which we call B_r^- , must remain constant:

$$\int_{\Omega} B_r^+ d\Omega = \int_{\Omega} B_r^- d\Omega = N/2 = \text{constant}$$

We can define the general radial core field B_r in terms of the spatially-varying intensities B_r^+ and B_r^- as $B_r = B_r^+ - B_r^-$. The problem is to reconstruct the two positive intensities B_r^+ and B_r^- from the surface data. For fixed total fluxes this is essentially a problem of combinatorics which has received much attention⁹⁻¹¹; the answer to this problem, the Maximum Entropy method, has had considerable success in astronomy and medical physics fields and this is the approach that I adopt here.

Properties of this method are well-known; it is frequently used in radio-astronomy to reconstruct images with high dynamic ranges. The results from certain dynamo simulations¹² show that it is possible for dynamo action to concentrate magnetic field into localized flux lines or 'ropes' in which B_r is strong and relatively weak elsewhere; this is a scenario where conventional quadratic regularization methods will produce poor images, because they bias strong localizations heavily towards zero. A method based on maximizing the entropy S of both B_r^+ and B_r^- where:

$$S(x_i) = \sum_i x_i - m_i - x_i \log(x_i/m_i)$$

and where x_i is an intensity in either B_r^+ or B_r^- with 'default' value m_i is much less likely to do this; indeed, synthetic tests on radial magnetic fields from numerical dynamo models confirm this property.

I apply the method to two high-quality data sets. The first is a selection of Magsat data from 1980, the first three-component magnetometry mission. The second is from the satellite Oersted, which was launched in February 1999 and is still collecting data; the selection of data is from December 1999–January 2000. The representation of the core magnetic field is in the form of the "spherical triangle tessellation"¹³ whereby the core is tessellated into $P = 1,442$ almost equally-spaced nodes and 2,880 spherical triangles. The node structure is based on the subdivision of the regular icosahedron¹⁴. Each of the P nodes specifies the core magnetic field by two positive numbers b_i^+ and b_i^- . Making the assumption that the mantle is an insulator, a valid assumption considering the frequencies present in the core spectrum, the forward problem is a convolution of the core field with a known kernel¹³. This kernel, the equivalent of the point-spread function in astronomical imaging, is very wide and therefore the deconvolution problem is quite severe.

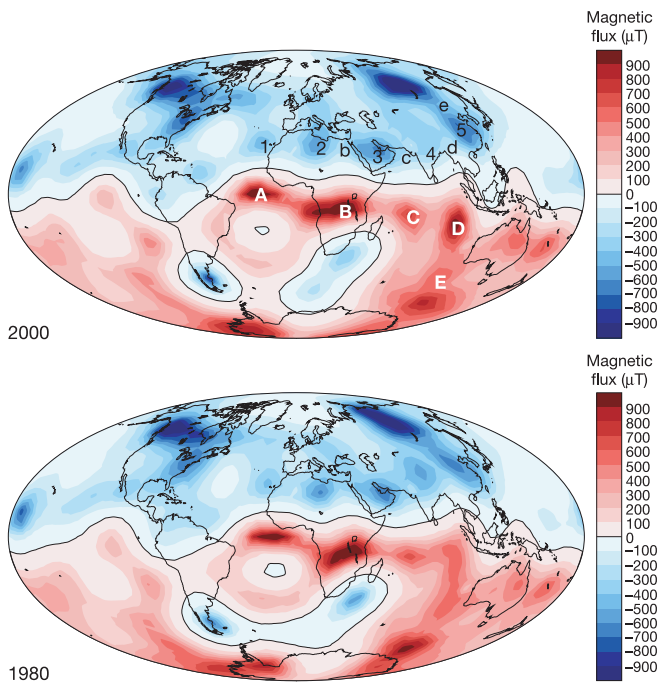


Figure 1 Comparison of the radial magnetic field for epochs 1980 and 2000 on Aitoff equal-area projection. Red colours represent magnetic flux out of the core, while blue colours represent magnetic flux entering the core; each colour bar represents an interval of 100 μT . The continental outlines are for orientation. The labels A–D and 1–5 define the flux spots referred to in the text; the high-latitude spots are E and e. In calculating the models the Magsat data are assigned errors of 10 nT, commensurate with previous studies. The Oersted data have lower errors and are treated using the anisotropic error model of Holme and Bloxham²¹; we use $\sigma = 2.25$ nT, and errors of 10° and 60° for the two pointing angle errors ψ and χ (ref. 22). The sensitivity of the data to the model is computed using the analytic solutions to the Neumann problem for Laplace's equation and a numerical integration over the five or six triangles adjacent to a node that have non-zero contribution to the sensitivity matrix¹³. The unsigned flux of the two models differs by less than 0.5%.

Table 1 Model comparison

Model	No. of data	Misfit = $\sqrt{\chi^2/N}$	α	S (μT)
1980	1,600 (Z)	1.00	3.82×10^{-2}	-873×10^3
2000	3,684 (X,Y,Z)	1.00	2.50×10^{-2}	-865×10^3

Data, misfit, trade-off parameter α and entropy S for the 1980 and 2000 models. X, Y, Z are the north, east and down components of the magnetic field. A 'default' value of $m_i = 10 \mu\text{T}$ has been used; this small value does not unduly penalize large-amplitude features.

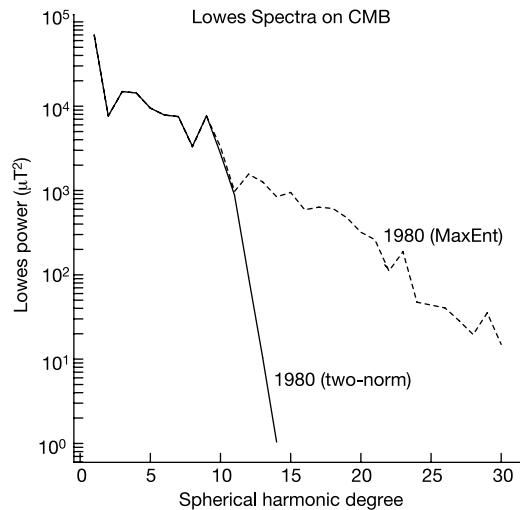


Figure 2 Comparison of the spherical harmonic energy spectrum (the contribution by spherical harmonic degree to $|\mathbf{B}|^2$) on the core surface for the conventional two-norm model and the maximum entropy model.

The solutions I compute assume gaussian errors in the data and therefore use a χ^2 measure of misfit. I aim to fit χ^2/N to unity with N data, and this is easily achieved, as shown in Table 1. The calculations use results from ref. 15 for the entropy and minimize:

$$\chi^2 - \alpha[S(b_i^+) + S(b_i^-)]$$

with α chosen so as to achieve the required data fit.

Figure 2 shows the spherical harmonic energy spectrum of these images (often called the Lowes spectrum). These spectra are much more realistic when compared to the output of numerical dynamo models⁷, which show a very slow decrease of magnetic energy with increasing spherical harmonic degree.

Figure 1 shows the fields on the core surface separated by 20 years, illustrating the intense equatorial spots on the core surface; similar modulations in field strength are evident, though not as clearly, in ref. 16. Both in the Northern and Southern hemispheres a high-amplitude strip of the field appears to be connected to the high-latitude flux spots *E*, *e* at longitude 120° E, and the azimuthal orientation of the modulations may be suggestive of some type of wave train.

Despite being based on different data sets the images are very coherent, and display many of the same features, though with slight translations, as one would expect for such a small difference in time. The spot currently centred over Africa has been known to exist as far back as 1590, slowly drifting to its current position through time from the Indian Ocean. Evidence suggesting the prolonged existence of these features is presented in Fig. 3 which shows a similar spot in a map of the time-averaged palaeomagnetic field¹⁷. We must ask ourselves why the geodynamo has such intense fields close to the Equator, a counter-intuitive result.

I speculate that these features could be the result of equatorially trapped waves belonging to Zhang's¹⁸ subclass of the eigenfunctions of the Poincaré equation, having strong alignment along the rotation axis, but whose phase speeds have been modified dramatically by an underlying toroidal magnetic field. I envisage the radial magnetic field to be passively modulated by the fluid motion, such that radial flux maxima coincide with regions of fluid convergence (that is, downwelling) and minima coincide with regions of fluid divergence. In a periodic wave motion the flux spots mark the neutral points in the tangential motion.

It is unlikely that we would observe a pure eigenmode from Zhang's theory¹⁸. However, I can illustrate the possibilities of the

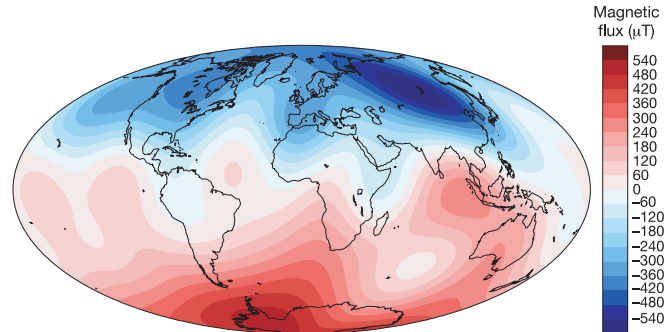


Figure 3 The palaeomagnetic time-average radial magnetic field for the last 2.5 Myr (ref. 17), derived by a conventional quadratic regularization method, showing the presence of an equatorial flux spot under the Indian Ocean. Each colour bar represents an interval of 60 μT ; Aitoff equal-area projection.

theory. In Fig. 1a I suspect that the phase difference between maxima B, C D and 2, 3, 4, 5 is significant; maxima in B, C and D coincide with minima at b, c and d. (The pair of spots A–I appear to be different, given their remarkable symmetry). Concentrations B, C and D are consistent with equatorially antisymmetric modes, and in terms of the simplest modes, we take $l - m = 3$, where l is the meridional and m the azimuthal wavenumber respectively. (Modes with $l - m = 1$ have purely toroidal velocity fields and would have no visible effect on B_r). These waves live in a 'equatorial waveguide' of width $\sqrt{2/m}$ and are members of the preferred convective instability at low Prandtl number; for $m \approx 8$, appropriate for the equatorial spots on Fig. 1, maxima are predicted at latitudes $\pm 20^\circ$, not dissimilar to the observed latitudes. In the nonmagnetic case these waves have extremely fast propagation speeds. However, certain idealized magnetic fields¹⁹ can lower the frequency of these waves from close to the diurnal frequency Ω to decade-to-century timescales.

For the simple magnetic field configuration amenable to analytical treatment corresponding to a constant current density through the core, the eigenmodes remain in the form of Poincaré modes¹⁹; these have frequencies ω (positive corresponding to westward propagation) of:

$$\omega = -\frac{mB^2}{\rho\mu_0 a^2 \Omega} \left[\frac{m - \lambda}{\lambda} \right]$$

in the presence of a toroidal magnetic field of size B , where ρ is the density, μ_0 the permeability of free space, a the radius of the core and λ is an eigenvalue of the Poincaré equation. At least one of these waves propagates westward: the eigenvalue $\lambda = -0.57$ for $m = 8$, $l = 11$ (ref. 18), giving timescale $T = 2\pi/\omega = 1,000$ years in the presence of a toroidal field of size 4.3 mT. This field strength is only three times the highest amplitudes on Fig. 1, and an order of magnitude stronger than the typical root-mean-square radial field strength; these figures would certainly not be considered unreasonable for the toroidal field strength.

Further analysis would take into account mode mixing and attenuation, and if continuous satellite monitoring of the core field were achievable on a multi-decadal timescale, it would be of interest to try to analyse the dispersion relations²⁰ of these features. Nevertheless, successful detailed simulation and comparison may verify the correspondence between the observations and theory, may explain the apparent preferential antisymmetric/westward-propagating mode selection and lead to useful constraints on the hidden toroidal field strength in the core. □

Received 14 January; accepted 24 June 2003; doi:10.1038/nature01879.

- Backus, G. E., Parker, R. & Constable, C. *Foundations of Geomagnetism* (Cambridge Univ. Press, Cambridge, 1996).
- Bloxham, J. & Gubbins, D. The secular variation of Earth's magnetic field. *Nature* **317**, 777–781 (1985).

3. Jackson, A., Jonkers, A. R. T. & Walker, M. Four centuries of geomagnetic secular variation from historical records. *Phil. Trans. R. Soc. Lond. A* **358**(1768), 957–990 (2000).
4. Backus, G. E. & Gilbert, F. Numerical applications of a formalism for geophysical inverse problems. *Geophys. J. R. Astron. Soc.* **13**, 247–276 (1967).
5. Parker, R. L. *Geophysical Inverse Theory* (Princeton Univ. Press, Princeton, 1994).
6. Tikhonov, A. N. & Arsenin, V. Y. *Solutions of Ill-posed Problems* (John Wiley and Sons, London, 1977).
7. Roberts, P. H. & Glatzmaier, G. A. A test of the frozen-flux approximation using a new geodynamo model. *Phil. Trans. R. Soc. Lond.* **358**, 1109–1121 (2000).
8. Bondi, H. & Gold, T. On the generation of magnetism by fluid motion. *Mon. Not. R. Astron. Soc.* **110**, 607–611 (1950).
9. Gull, S. F. & Daniell, G. J. Image reconstruction from incomplete and noisy data. *Nature* **272**, 686–690 (1978).
10. Gull, S. F. & Skilling, J. Maximum entropy method in image processing. *IEEE Proc.* **131**(F), 646–659 (1984).
11. Buck, B. & Macaulay, V. A. (eds) *Maximum Entropy in Action* (Oxford Univ. Press, Oxford, 1991).
12. Christensen, U., Olsen, P. & Glatzmaier, G. A. Numerical modelling of the geodynamo: a systematic parameter study. *Geophys. J. Int.* **138**, 393–409 (1999).
13. Constable, C. G., Parker, R. L. & Stark, P. B. Geomagnetic field models incorporating frozen-flux constraints. *Geophys. J. Int.* **113**, 419–433 (1993).
14. Baumgardner, J. & Fredrickson, P. O. Icosahedral discretization of the two-sphere. *SIAM J. Num. Anal.* **22**, 1107–1115 (1985).
15. Gull, S. F. & Skilling, J. *The MEMSYS5 User's Manual* (Maximum Entropy Data Consultants Ltd, Royston, 1990); available at (http://www.maxent.co.uk/documents/MemSys5_manual.pdf).
16. Hulot, G., Eymin, C., Langlais, B., Manda, M. & Olsen, N. Small-scale structure of the geodynamo inferred from Oersted and Magsat data. *Nature* **416**, 620–623 (2002).
17. Gubbins, D. & Kelly, P. Persistent patterns in the geomagnetic field during the last 2.5 Myr. *Nature* **365**, 829–832 (1993).
18. Zhang, K. On equatorially trapped boundary inertial waves. *J. Fluid Mech.* **248**, 203–217 (1993).
19. Malkus, W. V. R. Hydromagnetic planetary waves. *J. Fluid Mech.* **28**, 793–802 (1967).
20. Hide, R. Free hydrodynamic oscillations of the Earth's core and the theory of the geomagnetic secular variation. *Phil. Trans. R. Soc. Lond.* **259**, 615–647 (1966).
21. Holme, R. & Bloxham, J. The treatment of attitude errors in satellite geomagnetic data. *Phys. Earth Planet. Int.* **98**, 221–233 (1996).
22. Olsen, N. & members of the Ørsted International Science Team. Ørsted Initial Field Model. *Geophys. Res. Lett.* **27**, 3607–3610 (2000).

Acknowledgements The provision of high-quality Oersted data is due to the efforts of the Oersted Project team, for which I am grateful. I thank C. Constable and R. Parker for their hospitality, discussions and sharing of software, and C. Finlay for comments.

Competing interests statement The author declares that he has no competing financial interests.

Correspondence and requests for materials should be addressed to A.J. (a.jackson@earth.leeds.ac.uk).

In situ experimental evidence of the fate of a phytodetritus pulse at the abyssal sea floor

U. Witte¹, F. Wenzhöfer^{1*}, S. Sommer², A. Boetius^{1*}, P. Heinz³, N. Aberle^{1*}, M. Sand¹, A. Cremer², W.-R. Abraham⁴, B. B. Jørgensen¹ & O. Pfannkuche²

¹Max-Planck-Institute for Marine Microbiology, Celsiusstrasse 1, 28359 Bremen, Germany

²GEOMAR Research Center, Wischhofstrasse 1-3, 24148 Kiel, Germany

³Institute of Geosciences, University of Tübingen, Sigwartstrasse 10, 72076 Tübingen, Germany

⁴Gesellschaft für Biotechnologische Forschung, Mascheroder Weg 1, 38124 Braunschweig, Germany

* Present addresses: Marine Biological Laboratory, University of Copenhagen, Strandpromenaden 5, DK-3000 Helsingør, Denmark (F.W.); Alfred Wegener Institute for Polar and Marine Research, Columbusstrasse, 27515 Bremerhaven, Germany (A.B.); Max-Planck-Institute for Limnology, Post Box 165, 24302 Plön, Germany (N.A.)

More than 50% of the Earth's surface is sea floor below 3,000 m of water. Most of this major reservoir in the global carbon cycle and final repository for anthropogenic wastes is characterized by severe food limitation. Phytodetritus is the major food source for abyssal benthic communities, and a large fraction of the annual food load can arrive in pulses within a few days^{1,2}. Owing to

logistical constraints, the available data concerning the fate of such a pulse are scattered^{3,4} and often contradictory^{5–10}, hampering global carbon modelling and anthropogenic impact assessments. We quantified (over a period of 2.5 to 23 days) the response of an abyssal benthic community to a phytodetritus pulse, on the basis of 11 *in situ* experiments. Here we report that, in contrast to previous hypotheses^{5–11}, the sediment community oxygen consumption doubled immediately, and that macrofauna were very important for initial carbon degradation. The retarded response of bacteria and Foraminifera, the restriction of microbial carbon degradation to the sediment surface, and the low total carbon turnover distinguish abyssal from continental-slope 'deep-sea' sediments.

The Porcupine abyssal plain (PAP) is remote from both the continental slope to the east and the Mid-Atlantic Ridge to the west, and thus is largely unaffected by slope processes¹². The mean annual vertical influx of particulate organic carbon (C_{org}) of approximately $1 \text{ g C m}^{-2} \text{ yr}^{-1}$ has a strong seasonal pattern, reaching a maximum in mid-summer¹³. Prior to the mid-summer maximum, we simulated a settling food pulse by injection of ^{13}C -labelled phytodetritus into the chambers of benthic landers and then followed the pathway of tracer- ^{13}C through the abyssal community. Sedimentary C_{org} was 0.4 wt%, oxygen penetration depth was 15 cm, and the background sediment community oxygen consumption (SCOC) was $0.44 \pm 0.04 \text{ mmol O}_2 \text{ m}^{-2} \text{ d}^{-1}$.

The change in SCOC due to the enrichment with particulate organic matter (POM) was very rapid: SCOC increased significantly within 2.5 d, and the elevated remineralization rates lasted throughout the experiment (Fig. 1a). Bacteria usually dominate benthic biomass in deep-sea sediments, and are thought to be the primary agents of benthic carbon remineralization^{14,15}. At PAP, total microbial biomass accounted for approximately 95% of benthic biomass, thus most of the increase in SCOC would probably be due to an enhanced microbial respiration. A typical bacterial extracellular enzyme for the degradation of phytodetritus carbohydrates is β -glucosidase. It generally has a low activity in deep-sea sediments, and is substrate inducible—for example, by the addition of plant detritus¹⁶. Compared to SCOC, however, the onset of bacterial production of hydrolytic enzymes was retarded: extracellular enzyme activity (EEA) of β -glucosidase was unchanged after 2.5 d but had doubled after 8 d (Fig. 1b), whereas it decreased in the controls. The incorporation of ^{13}C from phytodetritus into bacterial phospholipid-derived fatty acids (PLFA) required even more time: the bacterial biomarkers contained only a moderate amount of excess ^{13}C after 8 d, but the specific uptake of tracer increased significantly within the second and third weeks of the experiment (Fig. 1c). Bacteria had assimilated approximately 1 mg of ^{13}C -labelled diatom carbon after 8 d; after 23 d, bacterial assimilation had increased to 4 mg $^{13}\text{C}_{org}$ (Fig. 2). Total bacterial cell numbers (0–5 cm) did not increase significantly, but it is possible that net growth of the microbial community was low owing to grazing by larger organisms.

Other than bacteria, Foraminifera and multicellular organisms can ingest fresh phytodetritus directly. Although the ^{13}C pulse chase experiments cannot distinguish between uptake and assimilation for these organisms, the foraminiferal response was retarded and the response pattern very similar to the pattern of extracellular enzyme activity and assimilation of ^{13}C by bacteria: after 2.5 d, labelling of Foraminifera was high, but total uptake of ^{13}C was still low. The specific uptake increased between 8 d and 23 d, and dominated the carbon uptake by the benthic biota after 23 d (Figs 1d and 2) despite the low total foraminiferal biomass. A significant increase in Foraminifera biomass was observed after 23 d. In contrast, metazoa became labelled immediately with ^{13}C (Figs 1e, f and 2). Macrofauna, in particular, gained immediate access to the phytodetritus: after 2.5 d, isotope signatures revealed that 77% of the organisms

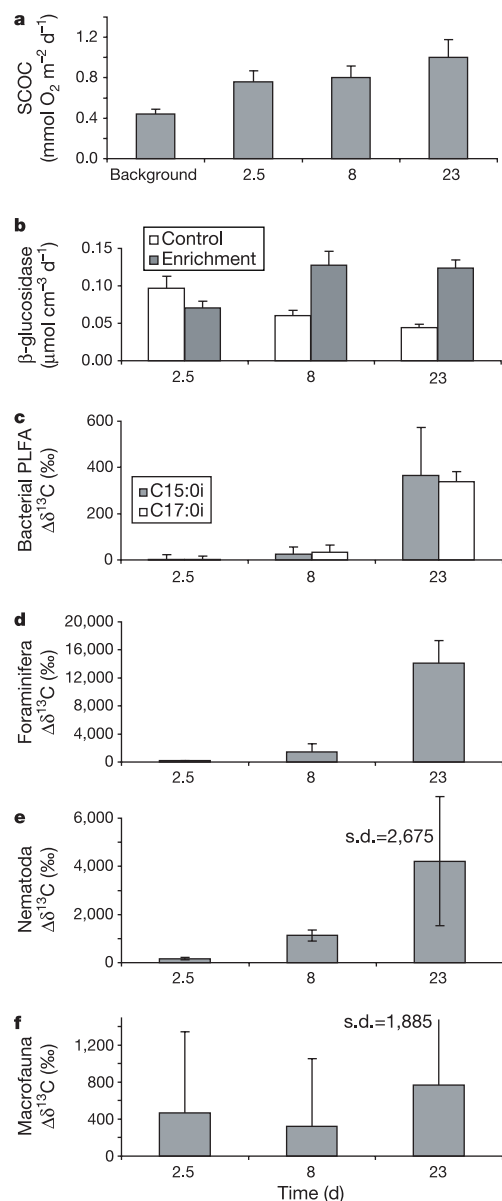


Figure 1 Response of different benthic compartments to POM enrichment. **a**, SCOC; **b**, extracellular enzyme activity of β -glucosidase; **c**, specific labelling patterns of bacterial PLFA ($\Delta\delta^{13}\text{C}$); **d–f**, mean specific uptake ($\Delta\delta^{13}\text{C}$) of Foraminifera (**d**, 0–1 cm), Nematoda (**e**, 0–5 cm) and macrofauna (**f**, 0–10 cm). Error bars represent standard deviations. High standard deviations for macrofauna and meiofauna are due to the strong individual differences in labelling between specimens (macrofauna labelling, for example, ranged from $\Delta\delta^{13}\text{C} = -20\text{‰}$ to $\Delta\delta^{13}\text{C} = +5,747\text{‰}$)²².

had ingested ^{13}C -labelled organic material. For total macrofauna, labelling was high throughout the experiments, but varied considerably between the different taxa¹⁷. The apparent decline of macrofaunal tracer uptake, from 3.4 mg after 8 d to 1 mg after 23 d, is probably the result of spatial variability in macrofaunal abundance¹⁷. Macrofauna in the 5–10-cm layer became labelled within days, and meiofauna (Nematoda) at 2–5 cm depth within weeks (Fig. 3a, b). In contrast, bacterial assimilation of organic carbon mixed below the surface layer was almost negligible: EEA of β -glucosidase at 2–5 cm had increased after 3 weeks (Fig. 3c), but only minute amounts of tracer were incorporated into subsurface bacterial biomass (Fig. 3d).

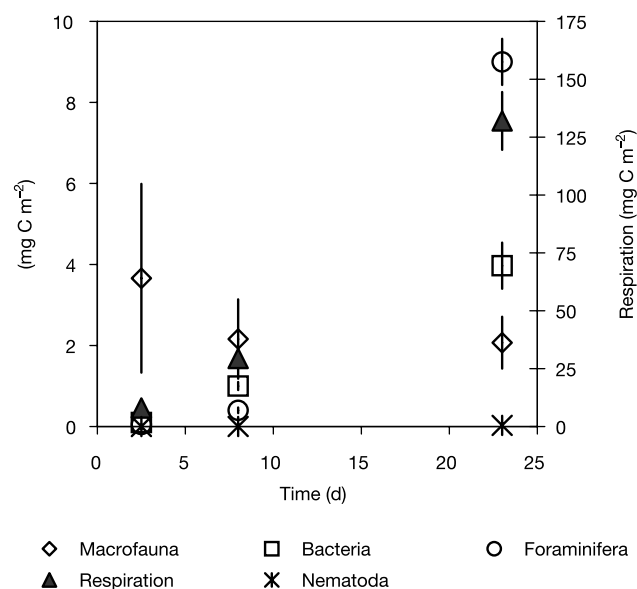


Figure 2 Pathways of ^{13}C labelled phytodetritus through the benthic community with time. (For comparison: total microbial biomass is approximately 2.5 g C m^{-2} ; macrofauna biomass, 120 mg C m^{-2} ; Foraminifera, 15 mg C m^{-2} ; meiofauna (Nematoda), 5 mg C m^{-2}).

In contrast to our experimental results, studies from the north-east Atlantic Ocean and the North Pacific Ocean reported a time lag of several weeks between the settling of a phytodetritus pulse and an increase in SCOC^{7,18,19}. This discrepancy may relate to the difficulty of following the response to a natural sedimentation event, as well as to the paucity of natural food prior to our experiments and the use of a ‘fresh’ monoculture different from the normal detrital food material. Despite their low significance in terms of biomass ($<5\%$), macrofauna initially dominated the benthic material processing. This implies that a large fraction of the material is passed through the gut system of a larger animal, where its composition is altered, before it becomes available for other (micro)organisms. Multi-cellular organisms initially outcompete bacteria owing to their higher capacity for directed movement, their ability to structure the sediment, and the advantage of internal digestion. These results are corroborated by the fact that the superabundance of megafauna at PAP could even prevent the accumulation of phytodetritus on the sea floor^{20,21}. However, bacteria still sustain a much greater biomass in abyssal sediments than metazoa because of their ability to degrade aged organic material as well as their capacity to survive starvation. At continental-slope depth, macrofauna rapidly subduct labile food down to 5–15 cm depth where it is degraded quickly by bacteria and deep-dwelling deposit feeders^{6,22,23}. In contrast, at abyssal depths, the absence of a deep-dwelling macrofaunal community and lower benthic biomass result in low total carbon processing rates that decrease with increasing depths from the continental shelf (61 mg C m^{-2} within 1.5 d; ref. 11) to slope ($9\text{--}25 \text{ mg C m}^{-2}$ within 1.5 d; refs 11, 23) to abyssal-plain sediments (6 mg C m^{-2} within 1.5 d). In our study, 14% of the 1 g C m^{-2} added were processed within 23 d. This suggests that a large phytodetritus pulse could sustain the elevated levels of activity for long periods (at least several months). This is long enough for many deep-sea organisms to complete their reproductive cycles, which might be synchronized with or triggered by sedimentation events^{4,24}.

The present experiments have successfully documented and quantified the first steps of early diagenesis in abyssal sediments.

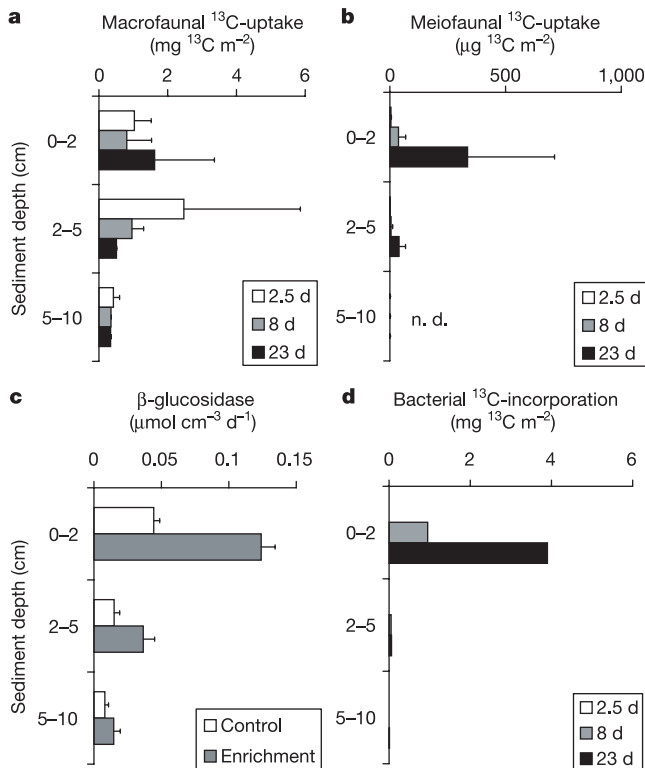


Figure 3 Vertical distribution of several quantities within the sediment. **a**, ^{13}C uptake by macrofauna; **b**, ^{13}C uptake by meiofauna (Nematoda); **c**, EEA of β -glucosidase; **d**, ^{13}C incorporation by bacteria. Error bars represent standard deviations.

It was demonstrated that there is no significant time lag between the arrival of a food pulse and the increase in SCOC. The macrofaunal community structure appears to be of greater importance for the initial stages of POM degradation. The bacterial and foraminiferal response was retarded, and microbial degradation of POM was restricted to the upper sediment horizon. The retarded response of microbiota, low rates of vertical mixing and the low carbon processing capacity underline the functional differences between continental-slope and deep-sea benthic communities. □

Methods

Experimental design

O_2 penetration depth was determined *in situ* with a profiling lander equipped with four O_2 microelectrodes and two deep-penetrating optical O_2 microsenors (optodes)²⁵. Enrichment experiments were carried out with three identical benthic chamber landers, each equipped with three chambers ($0.2 \times 0.2 \text{ m}$)²⁶. Each chamber of one lander carried two O_2 optodes that continuously recorded the O_2 concentration in the chamber water. To simulate a sedimentation event, a suspension of 0.2 g freeze-dried ^{13}C -labelled *Thalassiosira rotula* ($98 \pm 1\%$ ^{13}C ; $\text{C:N} = 13$), equivalent to $1 \text{ g C}_{\text{org}} \text{ m}^{-2}$, was injected into each chamber. During the long-term incubations, a pump system exchanged part of the chamber water each week. From each chamber, seven water samples (50 ml) were taken during the incubation, one before and after each pumping cycle. SCOC was determined by Winkler titration of syringe water samples and by optode readings, which agreed very well. Altogether, five short (2.5 d), three medium (8 d) and three long (23 d) deployments were carried out during RV *Poseidon* cruise 260 in May/June 2000 at $48^\circ 50' \text{ N}$; $16^\circ 35' \text{ W}$ ($4,800 \text{ m}$ water depth) in the PAP. In each lander, one chamber served as control without POM enrichment. Additional sediment cores were taken for the determination of background ^{13}C signatures of the organisms.

Sample processing

From each incubation, two subcores ($10 \times 20 \text{ cm}$ each) were sliced in layers of 0–1, 1–2, 2–5 and 5–10 cm and sieved over 250 and $63 \mu\text{m}$ for macrofauna and meiofauna (Nematoda). After taxonomic identification, all specimen were freeze-dried *in vacuo* for the subsequent determination of $^{12}\text{C}/^{13}\text{C}$ ratios with an ANCA 20–20 (Eurochem Scientific) isotope ratio mass spectrometer (IRMS)²⁵. For Foraminifera, $15\text{--}40 \text{ cm}^3$ of surficial sediment (0–1 cm) was sieved over $30 \mu\text{m}$ and stored frozen. Plasma-containing

Foraminifera were hand-picked and identified, then decalcified and dried prior to determination of $^{12}\text{C}/^{13}\text{C}$ ratios.

The remaining sediment was sliced in 0–2, 2–5, 5–10 cm, homogenized and sampled for bacterial biomass (cell counts), EEA of β -glucosidase and extraction of phospholipids^{27–29}. Concentrations of fatty acid methyl esters were determined by gas chromatography-flame ionization detection. Carbon isotope ratios as determined by GC-IRMS (Finnigan MAT 252 coupled with gas chromatograph combustion for compound identification) were corrected by using a mass balance for the one carbon atom in the methyl group added during derivatization.

The carbon isotope ratios are expressed in the delta notation ($\delta^{13}\text{C}$) relative to Vienna PDB: $\delta^{13}\text{C}(\text{‰}) = [(^{13}\text{C}/^{12}\text{C})_{\text{sample}} / (^{13}\text{C}/^{12}\text{C})_{\text{reference}} - 1] \times 1,000$. The uptake or incorporation of ^{13}C label by the organisms is expressed as specific uptake or excess ^{13}C above background (that is, $\Delta\delta^{13}\text{C} = \delta^{13}\text{C}_{\text{sample}} - \delta^{13}\text{C}_{\text{background}}$) or total uptake (I) (refs 23, 29). For bacteria, I was calculated from label incorporation into six bacterial PLFA (c15:0i, c16:0i, c16:1w7, c17:0i, c19:0i, c17:0; see refs 23, 29 and literature therein).

Received 24 January; accepted 2 June 2003; doi:10.1038/nature01799.

1. Billett, D. S. M., Lampitt, R. S., Rice, A. L. & Mantoura, R. F. C. Seasonal sedimentation of phytoplankton to the deep-sea benthos. *Nature* **302**, 520–522 (1983).
2. Lampitt, R. S. Evidence for the seasonal deposition of detritus to the deep-sea floor and its subsequent resuspension. *Deep-Sea Res.* **32**, 885–897 (1985).
3. Lochte, K. & Turley, C. M. Bacteria and cyanobacteria associated with phytodetritus in the deep-sea. *Nature* **333**, 67–69 (1988).
4. Tyler, P. A. Seasonality in the deep sea. *Oceanogr. Mar. Biol. Annu. Rev.* **26**, 227–258 (1988).
5. Smith, K. L. & Baldwin, R. J. Seasonal fluctuations in deep-sea sediment community oxygen consumption: Central and eastern north Pacific. *Nature* **307**, 624–626 (1984).
6. Graf, G. Benthic–pelagic coupling in a deep-sea benthic community. *Nature* **341**, 437–439 (1989).
7. Drzen, J. C., Baldwin, R. J. & Smith, K. L. Jr Sediment community response to a temporally varying food supply at an abyssal station in the NE Pacific. *Deep-Sea Res.* **45**, 893–913 (1998).
8. Sayles, F. L., Martin, W. R. & Deuser, W. G. Response of benthic oxygen demand to particulate organic carbon supply in the deep sea near Bermuda. *Nature* **371**, 686–689 (1994).
9. Smith, K. L. & Kaufmann, R. S. Long-term discrepancy between food supply and demand in the eastern North Pacific. *Science* **284**, 1174–1177 (1999).
10. Witbaard, R., Duineveld, G. C. A., Van der Weele, J. A., Berghuis, E. M. & Reyss, J. P. The benthic response to the seasonal deposition of phytopigments at the Porcupine Abyssal Plain in the North East Atlantic. *J. Sea Res.* **43**, 15–31 (2000).
11. Moodley, L. *et al.* Bacteria and Foraminifera: Key players in a short-term deep-sea benthic response to phytodetritus. *Mar. Ecol. Prog. Ser.* **236**, 23–29 (2002).
12. Billett, D. S. M. & Rice, A. L. The BENGAL programme: Introduction and overview. *Prog. Oceanogr.* **50**, 13–25 (2001).
13. Lampitt, R. S. *et al.* Material supply to the abyssal seafloor in the Northeast Atlantic. *Prog. Oceanogr.* **50**, 27–63 (2001).
14. Rowe, G. T. *et al.* ‘Total’ sediment biomass and preliminary estimates of organic carbon residence time in deep-sea benthos. *Mar. Ecol. Prog. Ser.* **79**, 99–114 (1991).
15. Turley, C. Bacteria in the cold deep-sea benthic boundary layer and sediment–water interface of the NE Atlantic. *FEMS Microbiol. Ecol.* **33**, 89–99 (2000).
16. Boetius, A. & Lochte, K. Effect of organic enrichments on hydrolytic potentials and growth of bacteria in deep-sea sediments. *Mar. Ecol. Prog. Ser.* **140**, 239–250 (1996).
17. Aberle, N. & Witte, U. Deep-sea macrofauna exposed to a simulated sedimentation event in the abyssal NE Atlantic: In situ pulse-chase experiments using ^{13}C -labelled phytodetritus. *Mar. Ecol. Prog. Ser.* **251**, 37–47 (2003).
18. Pfannkuche, O. Benthic response to the sedimentation of particulate organic matter at the BIOTRANS station, 47°N , 20°W . *Deep-Sea Res.* **40**, 135–149 (1993).
19. Smith, K. L., Baldwin, R. J., Karl, D. M. & Boetius, A. Benthic community responses to pulses in pelagic food supply: North Pacific Subtropical Gyre. *Deep-Sea Res.* **49**, 971–990 (2002).
20. Billett, D. S. M. *et al.* Long-term change in the megabenthos of the Porcupine Abyssal Plain (NE Atlantic). *Prog. Oceanogr.* **50**, 325–348 (2001).
21. Bett, B. J., Malone, M. G., Narayanaswamy, B. E. & Wigham, B. D. Temporal variability in phytodetritus and megabenthic activity at the seabed in the deep Northeast Atlantic. *Prog. Oceanogr.* **50**, 349–368 (2001).
22. Levin, L. A. *et al.* Rapid subduction of organic matter by maldanid polychaetes on the North Carolina slope. *J. Mar. Res.* **55**, 595–611 (1997).
23. Witte, U., Aberle, N., Sand, M. & Wenzhöfer, F. Rapid response of a deep-sea benthic community to POM enrichment: An *in situ* experimental study. *Mar. Ecol. Prog. Ser.* **251**, 27–36 (2003).
24. Goody, A. Biological responses to seasonally varying fluxes of organic matter to the ocean floor: A review. *J. Oceanogr.* **58**, 305–332 (2002).
25. Wenzhöfer, F., Holby, O. & Kohls, O. Deep penetrating benthic oxygen profiles measured *in situ* by oxygen optodes. *Deep-Sea Res.* **48**, 1741–1755 (2001).
26. Witte, U. & Pfannkuche, O. High rates of benthic carbon remineralisation in the abyssal Arabian Sea. *Deep-Sea Res.* **47**, 2785–2804 (2000).
27. Boetius, A., Ferdelmann, T. & Lochte, K. Bacterial activity in sediments of the deep Arabian Sea in relation to vertical flux. *Deep-Sea Res.* **47**, 2835–2875 (2000).
28. Boschker, H. T. S., de Brouwer, J. F. C. & Cappenberg, T. E. The contribution of macrophyte derived organic matter to microbial biomass in salt-marsh sediments: Stable isotope analysis of microbial biomarkers. *Limnol. Oceanogr.* **44**, 309–319 (1999).
29. Middelburg, J. J. *et al.* The fate of intertidal microphytobenthos carbon: An *in situ* ^{13}C -labelling study. *Limnol. Oceanogr.* **45**, 1224–1234 (2000).

Acknowledgements We thank the officers, crew and shipboard scientific party of RV *Poseidon* for their support at sea during the expedition POS 260/1-5, and M. Poser, W. Queisser, S. Meyer and V. Meyer for technical assistance with the lander systems. We also thank F. Lipschultz and U. Struck for help with isotope ratio measurements of macrofauna and Foraminifera, and K. Smith for review. This work was funded by the Bundesministerium für Bildung und Forschung as part of the BIGSET programme.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to U.W. (uwitte@mpi-bremen.de).

Climate change decreases aquatic ecosystem productivity of Lake Tanganyika, Africa

Catherine M. O'Reilly^{1*}, Simone R. Alin^{1*}, Pierre-Denis Plisnier², Andrew S. Cohen¹ & Brent A. McKee³

¹Department of Geosciences, University of Arizona, Tucson Arizona 85721, USA

²Royal Museum for Central Africa, 3080 Tervuren, and Department of Biology, Namur University, B-5000 Namur, Belgium

³Department of Earth and Environmental Sciences, Tulane University, New Orleans, Louisiana 70118, USA

* Present addresses: Environmental Science Program, Vassar College, Poughkeepsie, New York 12603, USA (C.M.O.); Large Lakes Observatory, University of Minnesota, Duluth, Minnesota 55812, USA (S.R.A.)

Although the effects of climate warming on the chemical and physical properties of lakes have been documented¹, biotic and ecosystem-scale responses to climate change have been only estimated or predicted by manipulations and models¹. Here we present evidence that climate warming is diminishing productivity in Lake Tanganyika, East Africa. This lake has historically supported a highly productive pelagic fishery that currently provides 25–40% of the animal protein supply for the populations of the surrounding countries². In parallel with regional warming patterns since the beginning of the twentieth century, a rise in surface-water temperature has increased the stability of the water column. A regional decrease in wind velocity has contributed to reduced mixing, decreasing deep-water nutrient upwelling and entrainment into surface waters. Carbon isotope records in sediment cores suggest that primary productivity may have decreased by about 20%, implying a roughly 30% decrease in fish yields. Our study provides evidence that the impact of regional effects of global climate change on aquatic ecosystem functions and services can be larger than that of local anthropogenic activity or overfishing.

Lake Tanganyika is a large (mean width, 50 km; mean length 650 km), deep (mean depth, 570 m; maximum depth, 1,470 m) north–south trending rift valley lake that is an important source of both nutrition and revenue to the bordering countries of Burundi, Tanzania, Zambia, and the Democratic Republic of Congo. The lake has historically supported one of the world's most productive pelagic fisheries³, and the annual harvest in recent years has been estimated to be between 165,000 and 200,000 metric tons (54–66 kg ha⁻¹), with an equivalent value of tens of millions of US dollars². The lake is oligotrophic and permanently thermally stratified with an anoxic hypolimnion. During the cool windy season (May to September), strong southerly winds tilt the thermocline, causing upwelling of deeper nutrient-rich waters at the south end of the lake and initiating seiche activity^{4,5}. Cooling during this season also contributes to a weaker thermocline, and entrainment of

deep nutrient-rich waters from the hypolimnion occurs in this time period⁴. Overall, these mixing events provide the dominant source of some limiting nutrients (P, Si) to the surface waters and are important in maintaining the pelagic food web^{4–6}.

Local records of air temperature from the Lake Tanganyika region show warming that is consistent with global patterns. Historical records show a rise of 0.5–0.7 °C in average annual air temperatures (Fig. 1a), consistent with the global increase of 0.6 ± 0.2 °C (ref. 7). The trend towards higher temperatures began primarily in the late 1970s and coincides with the timing of regional precipitation and temperature changes documented in several climate studies^{8,9}. On a regional scale, East Africa showed warm temperature anomalies from 1910 to 1930, from 1940 to 1960, and since the late 1970s (ref. 9).

The effects of climatic warming can be seen in water temperature data from Lake Tanganyika^{10–16}. Upper water-column temperatures (150-m depth) show a significant warming trend of 0.1 ± 0.01 °C per decade since 1913 ($r^2 = 0.76$, $F = 164$, $P < 0.0001$, $n = 53$, linear regression; Fig. 2a). Deep-water temperatures (600-m depth), which do not show seasonal variation and are generally homothermal between 400 m and 1,000 m, increased from 23.10 °C in 1938 to 23.41 °C in 2003 ($r^2 = 0.90$, $F = 74.3$, $P < 0.0001$, $n = 10$; refs 10–16 and Fig. 2b). This increase of 0.31 °C in deep-water temperature is comparable to that found in other African Great Lakes: Lake Victoria has warmed 0.3 °C between the 1960s and 1991 (ref. 17), the deep waters of Lake Malawi have warmed 0.29 °C since 1953 (refs 18, 19), and Lake Albert has warmed 0.5 °C since 1963 (ref. 20).

Along with increased temperatures, wind velocities in the Lake Tanganyika watershed have declined by 30% since the late 1970s (Fig. 1b). Records show that monthly averages of wind velocity during the cool windy season in the north remained constant at 2.2 ± 0.4 m s⁻¹ until 1985, after which they decreased significantly to 1.6 ± 0.3 m s⁻¹ (z-test, $Z = 7.54$, $P < 0.0001$, $n = 127$). In the

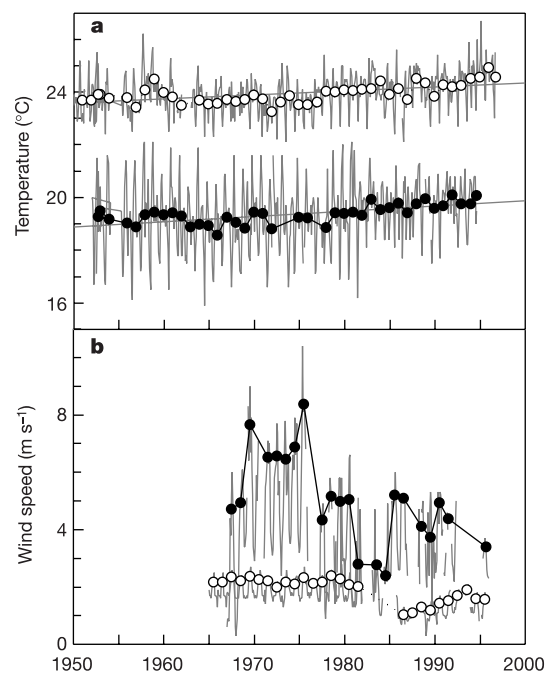


Figure 1 Historical meteorological records for the north (open circles) and south (filled circles) of Lake Tanganyika. **a**, Air temperatures. Monthly averages are shown in grey and are superimposed by annual means; regression (broken) lines are based on the full dataset. **b**, Wind speed. Monthly averages are shown in grey and are superimposed by windy season (May to September) means.

south, monthly averages for the cool windy season decreased significantly from $5.8 \pm 1.6 \text{ m s}^{-1}$ to $4.0 \pm 1.3 \text{ m s}^{-1}$ after 1977 (z -test, $Z = 7.38$, $P < 0.0001$, $n = 99$). Wind gusts were of short duration (hours) and were as high as 7.3 m s^{-1} , and velocities were dominated by the southern vector (data from May to September in 1993–1995 (ref. 21): $E = 1.3 \pm 0.3$, $W = 0.32 \pm 0.03$, $N = 0.7 \pm 0.2$, $S = 2.08 \pm 0.04 \text{ m s}^{-1}$).

The combined effect of increasing temperatures and decreasing wind speeds is to increase the stability of the lake and to reduce the mixing depth. We calculated that the stability of the water column during the non-windy season (defined as the work required to mix the water column to uniform density²²) increased by 97% from 84.4 kJ m^{-2} in 1913 to 166.3 kJ m^{-2} in 2003. Empirical evidence for reduced mixing is given by the depth of the oxygenated zone, which showed a significant shallowing trend of $1.3 \pm 0.2 \text{ m yr}^{-1}$ since 1939 to a current depth of around 80 m ($r^2 = 0.44$, $F = 27.1$, $P < 0.0001$, $n = 37$ linear regression; refs 11–16, 23, and Fig. 2c). If the 1938 measurements¹¹ are discarded because their determination involved visual colorimetric assessment (and is thus highly subject to bias), the trend is statistically stronger with a shallowing of $1.6 \pm 0.2 \text{ m yr}^{-1}$ ($r^2 = 0.67$, $F = 63.9$, $P < 0.0001$, $n = 34$). Together, these data suggest that increased thermal stability, coupled with a decline in wind velocity, has reduced mixing depth in the lake. This reduced mixing can be expected to diminish deep-water nutrient inputs to the surface waters, subsequently causing a decline in primary productivity rates.

A lake-wide decrease in productivity is consistent with trends seen in four sediment cores taken along 225 km of the eastern shoreline from watersheds with a range of land-use patterns (see Supplementary Information). Atomic C/N ratios of bulk organic

matter were low throughout all cores (grand mean 13.4 ± 0.3), indicating that organic matter was predominantly derived from autochthonous sources²⁴. Diagenetic alteration of organic material in these cores appeared to be negligible (see Supplementary Information). Carbon stable isotope records showed a trend towards more negative values beginning in the mid-1900s (Fig. 3). Carbon isotopes in sedimentary organic matter are widely regarded as indicators of phytoplankton productivity^{24–26}, and the sedimentary record of other lakes has been shown to record both isotope enrichment with increasing productivity and isotope depletion with decreasing productivity²⁷. Other possible explanations for a shift towards more negative carbon isotopes in Lake Tanganyika, such as the decrease in atmospheric isotope ratio $\delta^{13}\text{C}$ from fossil fuel combustion (the Suess effect), an increase in terrestrial organic matter, or a change in phytoplankton composition, can be discounted, and the shift occurs independently of the rates of sediment mass accumulation (see Supplementary Information).

There is evidence for rapid uptake of introduced deep-water nitrate after an upwelling event⁶, but phytoplankton do not seem to take up significant amounts of deep-water dissolved inorganic carbon (DIC), which has a $\delta^{13}\text{C}$ value that is 1.5‰ more negative than that of surface DIC¹⁵. Phytoplankton $\delta^{13}\text{C}$ values after an upwelling event were not significantly different from those before upwelling (t -test, $F = 0.25$, $P < 0.63$). Uptake of deep-water DIC may be limited because upwelling of nitrate (maximum concentration at a depth of 70–90 m)¹⁶ is not necessarily accompanied by significant upwelling of DIC (maximum concentration below a depth of 100 m)¹⁶. Generally, percentage carbon values in the sediment cores remain low (grand mean, $2.9 \pm 0.1\%$; see Supplementary Information), but three cores do indicate a decrease in carbon mass accumulation rates during the mid-1900s. Thus, paleolimnological indicators are consistent with a lake-wide decline in primary productivity over the past 80 years.

The timing of this shift towards more negative carbon isotope ratios coincides with changes in climatic temperature records, with the initial timing potentially obscured by human land use in the developed watersheds. The declining isotope trend began in the early 1900s immediately after the initial period of regional warming. The post-1950s trend followed the warm period of 1940, with an overall average decline slightly greater than $1.0\text{‰ } \delta^{13}\text{C}$. Assuming that the observed isotopic shift is wholly due to productivity changes, this would imply a roughly 20% drop in primary productivity rates²⁵. Previously established relationships for large, highly productive aquatic systems suggest that a 20% decrease in

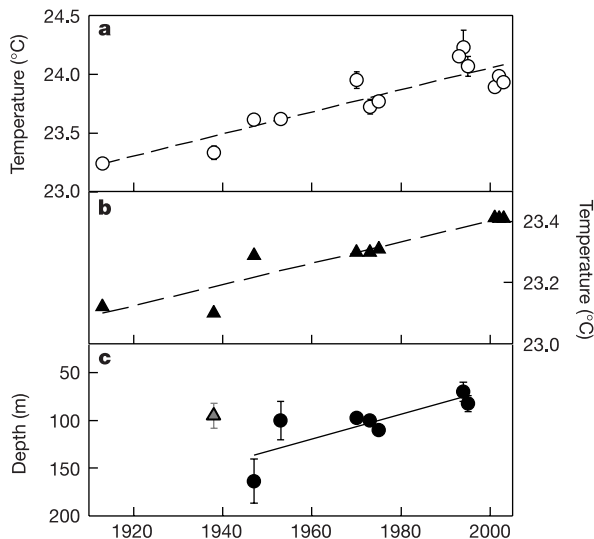


Figure 2 Historical and recent limnological measurements from the middle basin of Lake Tanganyika. **a**, Temperatures at 150 m in the non-windy warm season with linear regression line. Error bars represent the s.d., which provides an indication of intra-seasonal variability. Upper water temperature data for 150 m are from 1912 ($n = 1$)¹⁰, 1938 ($n = 3$)¹¹, 1946 ($n = 18$)^{12,13}, 1953 ($n = 3$)¹³, 1970 ($n = 2$)¹⁴, 1973 ($n = 11$)¹⁵, 1975 ($n = 1$)¹⁶, 1993 ($n = 2$), 1994 ($n = 3$), 1995 ($n = 3$), 2001 ($n = 1$, from the windy season), 2002 ($n = 3$, from the windy season) and 2003 ($n = 1$). **b**, Deep-water temperatures with linear regression line. Data shown are for 600 m and represent the relatively homothermal waters ($\pm 0.02^\circ\text{C}$) between 400 and 800 m (refs 10–16). **c**, Oxycline depth in the non-windy warm season. Oxycline data are from 1938 ($n = 3$)¹¹, 1946 ($n = 17$)^{12,13}, 1953 ($n = 3$)²³, 1970 ($n = 2$)¹⁴, 1973 ($n = 1$)¹⁵, 1975 ($n = 1$)¹⁶ and 1995 ($n = 10$). Grey triangle indicates the colorimetric visual measurements made in 1939 (ref. 11), which may be inaccurate because of methodology. The regression line does not include these data. Error bars represent the s.d.

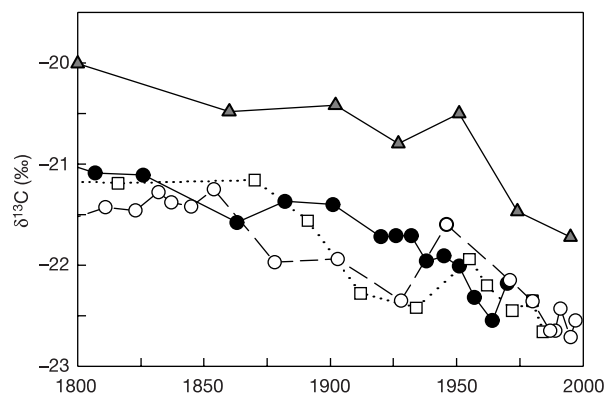


Figure 3 Carbon isotope records in sediment cores. $\delta^{13}\text{C}$ values indicate a post-1950s trend towards more negative values in all cores. Filled symbols represent cores from relatively undisturbed watersheds (filled circles, LT-98-58M; grey triangles, LT-97-56V), and open symbols represent cores from developed watersheds (squares, LT-98-37M; circles, LT-98-82M).

primary productivity would reduce fisheries yields by about 30% (ref. 3).

Finally, this inferred decrease in primary productivity helps to explain recent declines in the pelagic fishery that had been attributed to unknown environmental factors rather than to overfishing^{2,28–30}. The fishery is dominated by two clupeid species (*Stolothrissa tanganyicae* and *Limnothrissa miodon*) and a piscivore (*Lates stappersi*). The clupeid fishery is primarily supported by a cohort recruited during the cool windy season²⁹, and this seasonal pattern has been attributed to increased nutrient availability associated with upwelling⁴. Various studies have attributed the large (30–50%) decline in clupeid catch since the late 1970s partially to environmental factors because the lake had sustained high yields under similar fishing pressure for the previous 15–20 years (refs 2, 28–30). This decline was coincident with the disappearance of previously strong seasonal patterns in catch^{28,30}, suggesting a decoupling from ecosystem processes driven by the weakening of hydrodynamic patterns. The contribution (by weight) of the predator *Lates* species to the fish catch also declined from 20–60% beginning in 1955 to 2% after 1977 (ref. 30). These changes in the pelagic fishery are consistent with a lake-wide shift in ecosystem function.

The combined historical and paleolimnological data provide evidence that climate change has contributed to diminished productivity in Lake Tanganyika over the past 80 years. Within the next 80 years, air temperature increases of 1.3–1.7 °C are predicted for the Great Lakes region of East Africa⁹, which may further increase thermal stability and reduce productivity in these large lakes, provided that wind velocities remain low. The human implications of such subtle, but progressive, environmental changes are potentially dire in this densely populated region of the world, where large lakes are essential natural resources for regional economies. □

Methods

Data

Meteorological data from the north were for Bujumbura (3° 32' S, 29° 32' E, 772 m), obtained from the Institut Géographique du Burundi, and data from the south were for Mbala (8° 85' S, 31° 33' E, 1,632 m), obtained from the Department of Meteorology, Zambia.

Historical water temperature data were taken from the pelagic zone (>5 km offshore) of the middle basin of the lake near Kigoma, Tanzania, which has the most extensive historical record. Unless otherwise stated, all temperature and oxygen measurements were taken during the non-windy, warm season (October to April), when the lake is more stable. Early temperature measurements in 1913 and 1938 were determined using bottle samples rather than reversing thermometers, and we corrected these deep-water temperatures for adiabatic expansion. Stability of the water column was calculated with temperature profiles from the non-windy season to 700 m (ref. 22). Because salinity in Lake Tanganyika is low and we do not have historic salinity profiles, we assumed that salinity was negligible, which makes the stability values conservative estimates. The depth of the oxycline was conservatively defined as the last depth at which more than 0.5 mg l⁻¹ O₂ was measured.

Analysis of sediment cores

Cores LT-98-58M, LT-98-37M and LT-98-82M were dated using ²¹⁰Pb (see Supplementary Information). For these cores we analysed bulk organic matter for carbon isotopes roughly every 3 cm, with the exception of the top 3–6 cm, for which there was not enough material remaining after earlier work on some of these cores. Core LT-97-56V was dated using ¹⁴C (see Supplementary Information) and only the fine sediment fraction (<63 µm) from core intervals 0–1, 4–5, 8–9, 12–13, 16–17, 22–23 and 30–31 cm was analysed for ¹³C, as the fine sediment fraction should be influenced more directly by changes in autochthonous organic matter than should coarser sediment fractions³¹. The ^δ¹³C data from this core are potentially inconclusive, although they are consistent with data from the other cores.

All ¹³C samples were acidified with 10% HCl for at least 24 h and rinsed three times by decanting after centrifugation. Bulk sedimentary organic matter was analysed at the University of Waterloo Environmental Isotope Lab on an Isochrom continuous flow stable isotope mass spectrometer (Micromass) coupled to an elemental analyser CHNS-O EA1108 (Carla Erba) with a standard error of 0.05‰ for ^δ¹³C, 0.15‰ for ^δ¹⁵N, 0.02‰ for ‰C, and 0.3‰ for ‰N. The isotope ratios are expressed in delta notation (^δ¹³C = [(¹³C/¹²C)_{sample}/(¹³C/¹²C)_{standard}] – 1) with respect to deviation from a standard reference material (Pee Dee Belemnite).

Received 6 March; accepted 3 June 2003; doi:10.1038/nature01833.

1. McKnight, D., Brakke, D. F. & Mulholland, P. J. (eds) Freshwater ecosystems and climate change in North America. *Limnol. Oceanogr.* **41** (special issue) 815–1149 (1996).

- Molsa, H., Reynolds, J. E., Coenen, E. J. & Lindqvist, O. V. Fisheries research toward resource management on Lake Tanganyika. *Hydrobiologia* **407**, 1–24 (1999).
- Nixon, S. W. Physical energy inputs and the comparative ecology of lake and marine ecosystems. *Limnol. Oceanogr.* **33**, 1005–1025 (1988).
- Coulter, G. W. (ed.) *Lake Tanganyika and its Life* (Oxford Univ. Press, New York, 1991).
- Plisnier, P.-D. et al. Limnological annual cycle inferred from physical–chemical fluctuations at three stations of Lake Tanganyika. *Hydrobiologia* **407**, 45–58 (1999).
- O'Reilly, C. M., Hecky, R. E., Cohen, A. S. & Plisnier, P.-D. Interpreting stable isotopes in food webs: recognizing the role of time-averaging at different trophic levels. *Limnol. Oceanogr.* **47**, 306–309 (2002).
- Houghton, J. T. et al. in *Climate Change 2001: The Scientific Basis, Contribution of Working Group 1 to the Third Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge Univ. Press, Cambridge, 2001).
- Goddard, L. & Graham, N. E. Importance of the Indian Ocean for simulating rainfall anomalies over eastern and southern Africa. *J. Geophys. Res.* **104**, 19099–19116 (1999).
- Hulme, M., Doherty, R., Ngara, T., New, M. & Lister, D. African climate change: 1900–2100. *Climate Res.* **17**, 145–168 (2001).
- Marquardsen, H. Die Seen Tanganika. Moero und Bangweolo. *Mitteilungen aus den Deutschen Schutzgebieten Berlin* **29**, 97–98 (1916).
- Beauchamp, R. S. A. Hydrology of Lake Tanganyika. *Internationale Revue der Gesamten Hydrologie und Hydrographie* **39**, 316–353 (1939).
- Van Meel, L. *Contribution à la Limnologie de Quatre Grands Lacs du Zaïre Oriental: Tanganyika, Kivu, Mobutu Sese Seko (ex Albert), Idi Amin Dada (ex Edouard). Les Paramètres Chimiques, Fascicule A: Le Lac Tanganyika 1* (Institut Royal des Sciences Naturelles de Belgique, Brussels, Belgium, 1987).
- Van Meel, L. *Contribution à la Limnologie de Quatre Grands Lacs du Zaïre Oriental: Tanganyika, Kivu, Mobutu Sese Seko (ex Albert), Idi Amin Dada (ex Edouard). Les Paramètres Chimiques, Fascicule B: Le Lac Tanganyika 2* (Institut Royal des Sciences Naturelles de Belgique, Brussels, Belgium, 1987).
- Degens, E. T., von Herzen, R. P. & Wong, H.-K. Lake Tanganyika: water chemistry, sediments, geological structure. *Naturwissenschaften* **58**, 229–241 (1971).
- Craig, H. *Lake Tanganyika Geochemical and Hydrographic Study: 1973 Expedition* (Univ. California, San Diego, 1974).
- Edmond, J. M. et al. Nutrient chemistry of the water column of Lake Tanganyika. *Limnol. Oceanogr.* **38**, 725–738 (1993).
- Hecky, R. E. et al. Deoxygenation of the deep-water of Lake Victoria, East-Africa. *Limnol. Oceanogr.* **39**, 1476–1481 (1994).
- Beauchamp, R. S. A. Hydrological data from Lake Nyasa. *J. Ecol.* **41**, 226–239 (1953).
- Patterson, G. & Kachinjika, O. in *The Fishery Potential and Productivity of the Pelagic Zone of Lake Malawi* (ed. Menz, A.) 1–68 (Natural Resources Institute, Chatham, UK, 1995).
- Lehman, J. T., Litt, A. H., Mugidde, R. & Lehman, D. A. in *Environmental Change and Response in East African Lakes* (ed. Lehman, J. T.) 157–172 (Amsterdam, Kluwer, 1998).
- Verburg, P. et al. *Hydrodynamics of Lake Tanganyika and Meteorological Results* (FAO/FINNIDA Research for the Tanganyika Management of Fisheries on Lake Tanganyika, Bujumbura, Burundi, 1997).
- Idso, S. B. On the concept of lake stability. *Limnol. Oceanogr.* **18**, 681–683 (1973).
- Dubois, J. T. Evolution de la température, de l'oxygène dissous et de la transparence dans la baie nord du Lac Tanganyika. *Hydrobiologia* **10**, 215–240 (1958).
- Meyers, P. A. & Ishiwatari, R. Lacustrine organic geochemistry—an overview of indicators of organic matter sources and diagenesis in lake sediments. *Org. Geochem.* **20**, 867–900 (1993).
- Popp, B. N. et al. Effect of phytoplankton cell geometry on carbon isotopic fractionation. *Geochim. Cosmochim. Acta* **62**, 69–77 (1998).
- Hodell, D. A. & Schelske, C. L. Production, sedimentation, and isotopic composition of organic matter in Lake Ontario. *Limnol. Oceanogr.* **43**, 200–214 (1998).
- Brenner, M., Whitmore, T. J., Curtis, J. H., Hodell, D. A. & Schelske, C. L. Stable isotope (^δ¹³C and ^δ¹⁵N) signatures of sedimented organic matter as indicators of historic lake trophic state. *J. Paleolimnol.* **22**, 205–221 (1999).
- Shirakihara, K., Use, K., Kamikawa, S. & Mambona, W. B. Population changes of sardines in northern Lake Tanganyika. *African Study Monogr.* **13**, 57–67 (1992).
- Mannini, P. *Geographical Distribution Patterns of Pelagic Fish and Macrozooplankton in Lake Tanganyika* (FAO/FINNIDA Research for the Management of Fisheries on Lake Tanganyika, Bujumbura, Burundi, 1998).
- van Zwielen, P. A. M., Roest, F. C., Machiels, M. A. M. & van Densen, W. L. T. Effects of inter-annual variability, seasonality and persistence on the perception of long-term trends in catch rates of the industrial pelagic purse-seine fishery of northern Lake Tanganyika (Burundi). *Fisheries Res.* **54**, 329–348 (2002).
- Talbot, M. R. & Johannessen, T. A high resolution palaeoclimatic record for the last 27,500 years in tropical West Africa from the carbon and nitrogen isotope composition of lacustrine organic matter. *Earth Planet. Sci. Lett.* **110**, 23–37 (1992).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank M. O'Reilly for assistance with stability calculations, and M. Brenner, P. Brooks, and J. McManus for comments on the manuscript. The Tanzanian Commission for Science and the Tanzanian Immigration Agency provided research permits, and the Tanzania Fisheries Research Institute and the crew of the *R/V Explorer* provided field assistance. This work was funded by the National Science Foundation, Food and Agriculture Organization (FAO)/FINNIDA Research for the Management of the Fisheries of Lake Tanganyika Project, United Nations/Global Environment Facility Lake Tanganyika Biodiversity Project, University of Arizona Graduate College, Sigma Xi, Geological Society of America, and Office for Scientific, Technical and Cultural Affairs of the Prime Minister of Belgium.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.M.O. (caoreilly@vassar.edu).

Action plans used in action observation

J. Randall Flanagan¹ & Roland S. Johansson²

¹Department of Psychology and Centre for Neuroscience Studies, Queen's University, Kingston, Ontario K7L 3N6, Canada

²Section for Physiology, Department of Integrative Medical Biology, Umeå University, SE-90187 Umeå, Sweden

How do we understand the actions of others? According to the direct matching hypothesis, action understanding results from a mechanism that maps an observed action onto motor representations of that action^{1–4}. Although supported by neurophysiological^{1,5–13} and brain-imaging^{3,14–18} studies, direct evidence for this hypothesis is sparse. In visually guided actions, task-specific proactive eye movements are crucial for planning and control^{19–22}. Because the eyes are free to move when observing such actions, the direct matching hypothesis predicts that subjects should produce eye movements similar to those produced when they perform the tasks. If an observer analyses action through purely visual means, however, eye movements will be linked reactively to the observed action. Here we show that when subjects observe a block stacking task, the coordination between their gaze and the actor's hand is predictive, rather than reactive, and is highly similar to the gaze–hand coordination when they perform the task themselves. These results indicate that during action observation subjects implement eye motor programs directed by motor representations of manual actions and thus provide strong evidence for the direct matching hypothesis.

We tested the hypothesis that patterns of eye–hand coordination are similar when performing and observing a block stacking task. Previous studies of similar tasks have shown that there is a robust coupling between gaze and hand movements. Gaze leads the hand to blocks to be grasped and landing sites where blocks will be subsequently placed, and is rarely directed to the moving hand or block^{20,22–24}. These eye movements support hand-movement planning and control²² and may be viewed as part and parcel of the overall motor program for the task¹⁹.

The task involved moving three wooden blocks in the coronal plane. The blocks were all 2 cm in height and depth but had widths of 2, 3 and 4 cm. The locations of the blocks, as viewed by the subject, are shown in Fig. 1a. Before stacking, the blocks were aligned end to end at the right edge of a horizontal work surface and the task was to stack them, from the widest to the narrowest, at the left edge of the work surface. The same group of subjects both performed the task and observed an actor performing the task in counter-balanced order, although no order effects were observed in any of our analyses. The experimenter demonstrated the task once to each subject before data collection. In the action observation task, the actor sat across from the subject. Both the actor and the subjects used the tips of the right index finger and thumb to grasp the objects. In both conditions, the task was performed at a preferred rate and then at a faster rate.

Figure 1a shows all of the fixation points, from all subjects and trials, recorded in the action task performed at the preferred rate. The median path of the distal pad of the index finger is also shown. In agreement with previous reports, almost all fixations were directed towards sites of contact^{20,22,24}. Thus, the fixations were directed towards the grasp sites of the blocks to be picked up and the landing sites where the blocks were subsequently placed. The spatial distribution of fixation points in the corresponding action observation task (Fig. 1b) was very similar to that observed in the action task. In both tasks, gaze was predominantly directed to contact sites and subjects did not fixate the moving block or hand.

In Fig. 1c and d, the spatiotemporal coordination between gaze and hand movements in action and action observation, respectively, is examined by plotting the horizontal (x) gaze and hand positions as a function of time. Each red trace represents the x position of gaze during a single fixation (which can vary over time), and all fixations from all subjects and trials are shown. The blue and black curves represent the median x positions of gaze and of the distal pad of index finger, respectively. The median vertical (y) position of the index finger is shown in Fig. 1e and f. To preserve phase information when combining data from different trials, we first segmented each trial into phases and then normalized the time base of each phase to the median duration of that phase²².

In the action task (Fig. 1c, e), subjects fixated each forthcoming grasp and landing site well before the index finger arrived within the vicinity of the site. By contrast, gaze exited the grasp and landing sites at about the same time as the index finger exited these contact sites. This pattern of eye–hand coordination matches our previous work on object manipulation²². A similar pattern was observed in the action observation task (Fig. 1d, f). The sole exception was the first grasp site. In the action observation task, gaze exited the first grasp site slightly after the hand started to move away with the block. Although the lead of gaze over the hand appeared greater in the action task, in both tasks the gaze clearly anticipated forthcoming grasp and landing sites.

To quantify the coordination between eye and hand movements, we determined for each trial the times at which gaze and the index finger entered and exited each grasp and landing site zone. We defined the centre of each grasp zone as the position of the index finger when it reached a minimum in the vertical while grasping. The same method was used to define the centre of each landing site, except that we added a 1-cm downward offset to the measured minimum position of the index finger because the landing surface

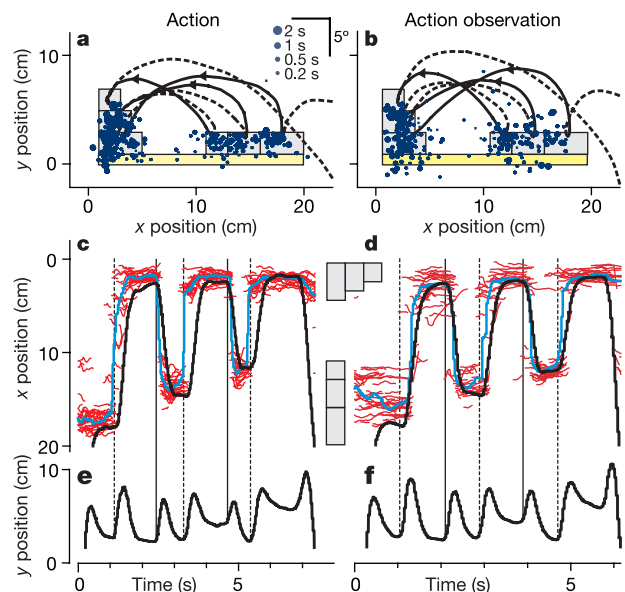


Figure 1 Gaze–hand coordination in action and action observation. **a, b**, Gaze positions at the end of periods between saccades (blue circles scaled to fixation duration), median hand path with and without a block in hand (unbroken and broken black lines, respectively), and approximate block positions before (right) and after (left) stacking. **c, d**, Median horizontal (x) positions of gaze (blue) and the index finger (black) as a function of time. Red traces represent the x position of gaze for all periods between saccades. Approximate x locations of the blocks are shown. **e, f**, Median vertical (y) position of the index finger. **c–f**, Time normalized to the median duration of all trials (Methods). Broken and unbroken vertical lines indicate the times at which the index finger exited grasp and landing sites, respectively.

was about 1 cm below the fingertip when placing blocks. Gaze and the index finger were deemed to be inside a given contact zone when they were within 2 cm (3°) and 0.5 cm (0.75°) of the centre position, respectively. The 3° radius for gaze was selected as it represents the size of the functional fovea in natural manipulation tasks²².

We analysed data from all contact sites with two exceptions. For gaze entries we excluded the first grasp site because gaze could arrive at this site well before the subject initiated the stacking task, and for gaze exits we excluded the last landing site because gaze could remain there until well after the task was completed. In both action and action observation, gaze arrived at a given contact zone before the hand, and the lead of gaze entry over hand entry increased with trial duration.

To assess the dependence of gaze entry times on trial duration, task and the interaction between duration and task, we carried out a stepwise linear regression analysis. The best regression model was one that included separate intercepts for the action and action observation tasks ($P < 0.001$), but had a common slope (-0.082 , $P < 0.001$) relating entry time to trial duration. On average, gaze arrived 150 ms earlier, relative to the index finger, in the action task. Unlike the gaze entry times, the gaze exit times did not vary with trial duration in either task, but stepwise linear regression analysis indicated separate intercepts for action and action observation ($P < 0.001$). On average, gaze departed the contact zone 72 ms before the index finger in the action task and 38 ms after the index finger in the action observation task. In both action and action observation, however, gaze shifts away from the contact zones were anticipatory or proactive. Had such gaze shifts been triggered by visual information related to hand movement away from the contact zone, we would expect gaze exits to lag behind hand exits by at least 100 ms (ref. 25).

To assess whether proactive eye movements require observing an interaction between the hand of the agent of the action and an object¹⁻⁴, we tested an additional group of subjects who observed the block stacking task being performed by the actor, no part of whom could be seen by the subjects (see Methods). Although subjects also

directed gaze to the block to be grasped and landing sites in this task (Fig. 2a), the pattern of eye-hand coordination was very different to that seen in the action and action observation tasks.

First, eye movements were linked reactively, rather than proactively, to the course of events. Gaze did not arrive at forthcoming block landing sites ahead of the block (Fig. 2b), and gaze exits from the grasp sites were markedly delayed (~ 200 ms) relative to the times at which blocks started to move (Fig. 2b and c, broken vertical lines). Second, fixations were distributed widely in the workspace (Fig. 2a); in fact, gaze frequently tracked the blocks when they moved to the stacking area. This is seen in Fig. 2b by the close match between the x position of the index finger when it moved to the landing site (unbroken black lines) and the x positions of gaze between saccades (red lines). Also note that the median x position of gaze (blue line) closely overlapped or lagged the x position of the index finger. Tracking was not observed when the actor's hand returned to pick up the next block, a movement that was invisible to the subject (Fig. 2b, c, broken segments of the black curves).

Figure 2d illustrates the tracking behaviour during block observation in the work plane by showing the gaze positions in the work plane between saccades. The longer red traces show periods of marked tracking, and the gaze position during these periods closely matched the path of the index finger (Fig. 2a). Figure 2e shows the corresponding plot for data combined from the action and action observation tasks, which were also performed at the preferred rate. Despite the fact that data from twice as many trials are shown, very few tracking periods can be observed.

To quantify the gaze tracking, we computed, for each subject, the total distance travelled by gaze (between saccades) that occurred during hand movements with the block, excluding periods when either the fingertip or gaze were within 2 cm (3°) of a grasp or landing site. We then expressed this as a percentage of the total distance travelled by the index finger with the block in hand and outside the contact zones. A between-subjects analysis of variance (ANOVA) indicated that the percentage of tracking was significantly greater ($F_{1,14} = 21.88$, $P < 0.001$) in the block observation task (mean $\pm$ s.d., 22.2 ± 10.8) than in the action and action observation tasks in which subjects viewed the hand (3.96 ± 1.9 ; data combined from both tasks).

The central advance of this study is the demonstration that during action observation humans implement, instinctively, motor programs equivalent to those used in action. Specifically, subjects activate highly similar eye motor programs when performing and observing the same task. In addition, these programs operate almost in phase with the corresponding programs run by the actor, because the eye-hand coordination was essentially the same in the two tasks. Thus, as postulated by the direct matching hypothesis^{4,26}, a part of the motor system of the observer resonated with that of the actor.

The tight linkage, in time and space, between actions of the hand in manipulation and eye movements²² implies that the control program for a particular action includes directions for the oculomotor and visual processing systems¹⁹. Therefore, the task-specific eye movements that we observed during action observation are probably linked to parts of the neural processes that account for planning and control of the manual action. We can thus infer that such processes function in action observation, which supports the hypothesis that action understanding is based on a direct matching mechanism that maps the visual representation of the observed action onto a motor representation of the same action.

By contrast, our results provide little support for the alternative view that the observer captures the actor's behaviour through purely visual analysis of the elements that form the action, such as moving objects or a moving hand (reviewed in refs 4, 27, 28). First, rather than being reactively linked to events of the scenery, gaze specifically predicted forthcoming contact events in action observation. This proactive gaze behaviour is in line with the event-based predictive

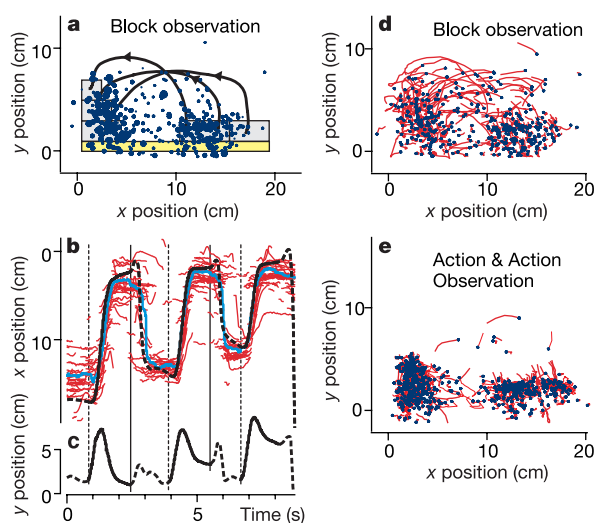


Figure 2 Gaze behaviour in the block observation task. **a–c**, Plots corresponding to those shown in Fig. 1; the unbroken black lines show the median path of the moving block and the broken segments in **b** and **c** mark periods when the actor's hand, invisible to the subjects, returned to pick up the next block. **d, e**, Gaze movements in the work plane during periods between saccades (red traces) for all subjects and trials in the block observation task (**d**) and for the action and action observation tasks combined (**e**). Blue circles indicate gaze positions at the start of each such period.

control strategies typical for manipulation²⁹. Second, in action observation, as in action, gaze was rarely directed towards the moving hand, as might be expected if hand motion were being visually analysed³⁰. Third, when subjects observed blocks moving without hand involvement, the gaze pattern differed from that engaged in action and action observation by being reactively coupled to the events. This result matches findings showing that the part of the motor representation to which the mirror system belongs is activated only when the observer views an object-oriented goal-directed task and not when the observer views its components^{1,2,18}. □

Methods

Subjects

Nine women and seven men aged from 19 to 30 years participated after providing written informed consent. All subjects were healthy, were right-handed, had normal vision and were naive as to the purpose of the experiment.

Block stacking task

The wooden blocks were located on a 19-cm wide work surface (Fig. 1a) centred 39 cm in front of the eyes and 8 cm below eye level. Eight subjects both performed the block stacking task and observed an actor performing the task who sat across from them. In both tasks, nine stacking trials performed at a preferred rate were followed by nine trials performed at a faster rate. In each block of nine trials, six were performed with an obstacle in the form of a vertically oriented block located in the centre of the workspace. These trials were interspersed among three trials without an obstacle. We focused our analysis on the latter trials. Each stacking trial was followed by an unstacking trial and between all trials subjects moved the hand to a parking position located 29 cm below and 4 cm to the right of the work surface.

Eight additional subjects observed the same block stacking task performed at the preferred rate by the same actor who was hidden from view. In this condition, the actor was positioned below the work surface and reached up to grasp tabs attached to the back surface of each block. The actor wore black gloves and a black outfit so as not to be visible against the black drape positioned behind the work surface in all conditions. Subjects were given no instructions on where to look in any of the three block stacking tasks.

Gaze position of the right eye in the plane in which the blocks were moved was recorded using an infrared eye tracking system and the position of the tip of the right index finger of the subject or actor were recorded with miniature electromagnetic sensors attached to the nail. A haptic calibration procedure was used to estimate the position on the distal pad of the index finger when in contact with a block. A bite bar was used to stabilize the head. The apparatus, calibration procedures and the accuracy and resolution of all measures have been described in detail elsewhere²².

Data processing

To preserve phase information when combining data from different trials for plotting, we first segmented each trial into seven contiguous phases based on when the index finger crossed a vertical line located at $x = 6$ (see Fig. 1a); the start and end of hand movement defined the beginning and end of the first and last phases, respectively. We then normalized the time base of each phase to the median duration of that phase. We defined trial duration, for use in statistical analysis, as the time from the end of the first phase to the start of the last. We detected the occurrences of saccades based on a filter applied to the gaze position signals combined with a threshold criterion as described²².

Received 6 May; accepted 13 June 2003; doi:10.1038/nature01861.

- Gallese, V., Fadiga, L., Fogassi, L. & Rizzolatti, G. Action recognition in the premotor cortex. *Brain* **119**, 593–609 (1996).
- Rizzolatti, G., Fadiga, L., Gallese, V. & Fogassi, L. Premotor cortex and the recognition of motor actions. *Brain Res. Cogn. Brain Res.* **3**, 131–141 (1996).
- Iacoboni, M. *et al.* Cortical mechanisms of human imitation. *Science* **286**, 2526–2528 (1999).
- Rizzolatti, G., Fogassi, L. & Gallese, V. Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Rev. Neurosci.* **2**, 661–670 (2001).
- di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V. & Rizzolatti, G. Understanding motor events: a neurophysiological study. *Exp. Brain Res.* **91**, 176–180 (1992).
- Fadiga, L., Fogassi, L., Pavesi, G. & Rizzolatti, G. Motor facilitation during action observation: a magnetic stimulation study. *J. Neurophysiol.* **73**, 2608–2611 (1995).
- Hari, R. *et al.* Activation of human primary motor cortex during action observation: a neuromagnetic study. *Proc. Natl Acad. Sci. USA* **95**, 15061–15065 (1998).
- Nishitani, N. & Hari, R. Temporal dynamics of cortical representation for action. *Proc. Natl Acad. Sci. USA* **97**, 913–918 (2000).
- Strafella, A. P. & Paus, T. Modulation of cortical excitability during action observation: a transcranial magnetic stimulation study. *NeuroReport* **11**, 2289–2292 (2000).
- Gangitano, M., Mottaghy, F. M. & Pascual-Leone, A. Phase-specific modulation of cortical motor output during movement observation. *NeuroReport* **12**, 1489–1492 (2001).
- Umiltà, M. A. *et al.* I know what you are doing: a neurophysiological study. *Neuron* **31**, 155–165 (2001).
- Avikainen, S., Forss, N. & Hari, R. Modulated activation of the human SI and SII cortices during observation of hand actions. *NeuroImage* **15**, 640–646 (2002).
- Kohler, E. *et al.* Hearing sounds, understanding actions: action representation in mirror neurons. *Science* **297**, 846–848 (2002).
- Grafton, S. T., Arbib, M. A., Fadiga, L. & Rizzolatti, G. Localization of grasp representations in humans

- by positron emission tomography. 2. Observation compared with imagination. *Exp. Brain Res.* **112**, 103–111 (1996).
- Rizzolatti, G. *et al.* Localization of grasp representations in humans by PET: 1. Observation versus execution. *Exp. Brain Res.* **111**, 246–252 (1996).
- Decety, J. *et al.* Brain activity during observation of actions. Influence of action content and subject's strategy. *Brain* **120**, 1763–1777 (1997).
- Iacoboni, M. *et al.* Reafferent copies of imitated actions in the right superior temporal cortex. *Proc. Natl Acad. Sci. USA* **98**, 13995–13999 (2001).
- Koski, L. *et al.* Modulation of motor and premotor activity during imitation of target-directed actions. *Cereb. Cortex* **12**, 847–855 (2002).
- Land, M. F. & Furneaux, S. The knowledge base of the oculomotor system. *Phil. Trans. R. Soc. Lond. Biol.* **352**, 1231–1239 (1997).
- Land, M., Mennie, N. & Rusted, J. The roles of vision and eye movements in the control of activities of daily living. *Perception* **28**, 1311–1328 (1999).
- Land, M. F. & McLeod, P. From eye movements to actions: how batsmen hit the ball. *Nature Neurosci.* **3**, 1340–1345 (2000).
- Johansson, R. S., Westling, G., Bäckström, A. & Flanagan, J. R. Eye-hand coordination in object manipulation. *J. Neurosci.* **21**, 6917–6932 (2001).
- Ballard, D. H., Hayhoe, M. M., Li, F. & Whitehead, S. D. Hand-eye coordination during sequential tasks. *Phil. Trans. R. Soc. Lond. Biol.* **337**, 331–338 (1992).
- Ballard, D. H., Hayhoe, M. M. & Pelz, J. B. Memory representations in natural tasks. *J. Cogn. Neurosci.* **7**, 66–80 (1995).
- Becker, W., (ed. Carpenter, R. H. S.) 95–137 (Macmillan, London, 1991).
- Rizzolatti, G., Fadiga, L., Fogassi, L. & Gallese, V. Resonance behaviors and mirror neurons. *Arch. Ital. Biol.* **137**, 85–100 (1999).
- Perrett, D. I. *et al.* Frameworks of analysis for the neural representation of animate objects and actions. *J. Exp. Biol.* **6**, 87–113 (1989).
- Allison, T., Puce, A. & McCarthy, G. Social perception from visual cues: role of the STS region. *Trends Cogn. Sci.* **4**, 267–278 (2000).
- Johansson, R. S. & Cole, K. J. Sensory-motor coordination during grasping and manipulative actions. *Curr. Opin. Neurobiol.* **2**, 815–823 (1992).
- Mataric, M. J. & Pomplun, M. Fixation behavior in observation and imitation of human movement. *Brain Res. Cogn. Brain Res.* **7**, 191–202 (1998).

Acknowledgements We thank G. Westling and A. Bäckström for engineering and software support. This study was supported by the Canadian Institutes of Health Research, the Swedish Medical Research Council, the Göran Gustafsson Foundation for Research in Natural Sciences and Medicine, and the 5th Framework Program of European Union.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.F. (flanagan@psyc.queensu.ca).

Prediction of auditory spatial acuity from neural images on the owl's auditory space map

Avinash D. S. Bala, Matthew. W. Spitzer & Terry T. Takahashi

Institute of Neuroscience, University of Oregon, Eugene, Oregon 97403, USA

The owl can discriminate changes in the location of sound sources as small as 3° and can aim its head to within 2° of a source^{1,2}. A typical neuron in its midbrain space map has a spatial receptive field that spans 40°—a width that is many times the behavioural threshold³. Here we have quantitatively examined the relationship between neuronal activity and perceptual acuity in the auditory space map in the barn owl midbrain. By analysing changes in firing rate resulting from small changes of stimulus azimuth, we show that most neurons can reliably signal changes in source location that are smaller than the behavioural threshold. Each source is represented in the space map by a focus of activity in a population of neurons. Displacement of the source causes the pattern of activity in this population to change. We show that this change predicts the owl's ability to detect a change in source location.

We measured spatial discrimination behaviour by using the

pupillary dilation response (PDR)². The barn owl's pupil dilates on presentation of a sound and habituates on repeated presentation of the same stimulus. When the stimulus is changed by presenting the same sound from a different location, the PDR recovers (Fig. 1a). We habituated the PDR by repeated presentation of a broadband noise (Fig. 1a, black marker) from a source at azimuth x° , and looked for recovery on presenting the identical noise burst from another source at azimuth $x^\circ + \Delta x^\circ$. Discrimination was quantified by expressing the difference in the PDR magnitudes evoked by habituating stimuli (from azimuth -3° ; Fig. 1a, broken lines) and test stimuli (from azimuth 3° ; Fig. 1a, unbroken lines) in units of their combined s.d. A non-parametric index of discrimination, standard separation⁴, D , was computed as follows:

$$D = \left| (\mu_x - \mu_{x+\Delta x}) / \sqrt{(\sigma_x^2 + \sigma_{x+\Delta x}^2)} \right| \quad (1)$$

where μ_x and $\mu_{x+\Delta x}$ refer to the mean magnitudes of the PDR to habituating and test stimuli, and σ_x and $\sigma_{x+\Delta x}$ are the respective s.d. D does not saturate, unlike another non-parametric measure of discrimination, proportion correct, $p(c)$, which allowed us to assess the relationship between behavioural and neuronal discrimination above threshold.

In Fig. 1b, behavioural discrimination ($D_{\text{behaviour}}$) is plotted against the angular separation (Δx°) of the two sources for three birds. The minimum audible angle (MAA) was defined as the smallest angular separation giving rise to $D_{\text{behaviour}} \geq 0.8$, which corresponds to 0.75 $p(c)$. The MAA was 3° for all three birds tested. Identical MAAs were observed according to a t -test ($P < 0.01$). This MAA is within the range of head-aim errors (1.7° to 4°) in a head-pointing task¹, another measure of spatial acuity. The psychometric functions shown in Fig. 1b can also be thought of as "generalization gradients"⁵—that is, functions that describe the recovery of a habituated response as test and habituating stimuli become more dissimilar. Space-specific neurons (SSNs) in the owl midbrain, which constitute the space map, respond to source displacement

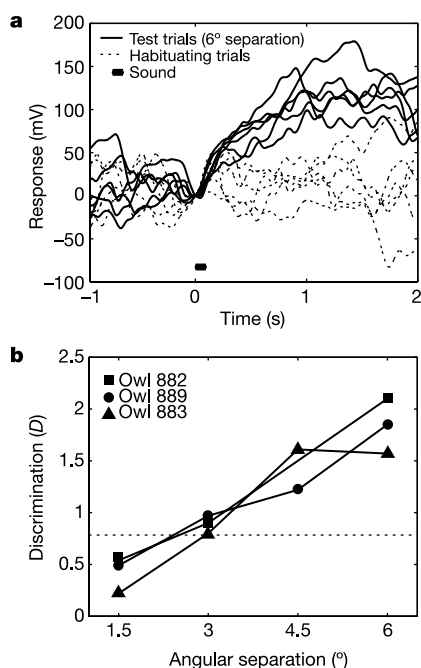


Figure 1 Behavioural discrimination of azimuthal source separation. **a**, Voltage traces representing pupillary responses to test (unbroken lines) and habituating (broken lines) stimuli from a single session (owl 883). Sources in this session were separated by 6° in azimuth. **b**, Behavioural discrimination measured in three subjects. Horizontal line represents the arbitrary discrimination threshold ($D = 0.8$).

by changing their firing rate. This change in firing rate provides a basis for recovery of the habituated response, even though SSNs themselves do not habituate. We investigated the relationship between the behavioural generalization gradient described above and the rate of change of neuronal activity in the space map.

The change in the neuronal image of the space map caused by displacing the source from x to $x + \Delta x$ was assessed in SSNs tuned to frontal space (-20° to $+20^\circ$ azimuth; -20° to $+20^\circ$ elevation; Fig. 2a). The spatial receptive field (SRF) of isolated SSNs was first charted coarsely using broadband noise presented in virtual auditory space (VAS⁶; Fig. 2b). A cross-section through the centre of the SRF was then sampled in 1° increments (Fig. 2b, white line, and 2c, black line). Neuronal discrimination (D_{neuron}) was then computed by assessing the change in firing rates evoked by sources at each pair of azimuthal locations separated by Δx° using equation (1). In the neuronal case, μ_x and $\mu_{x+\Delta x}$ refer to the mean firing rates at two

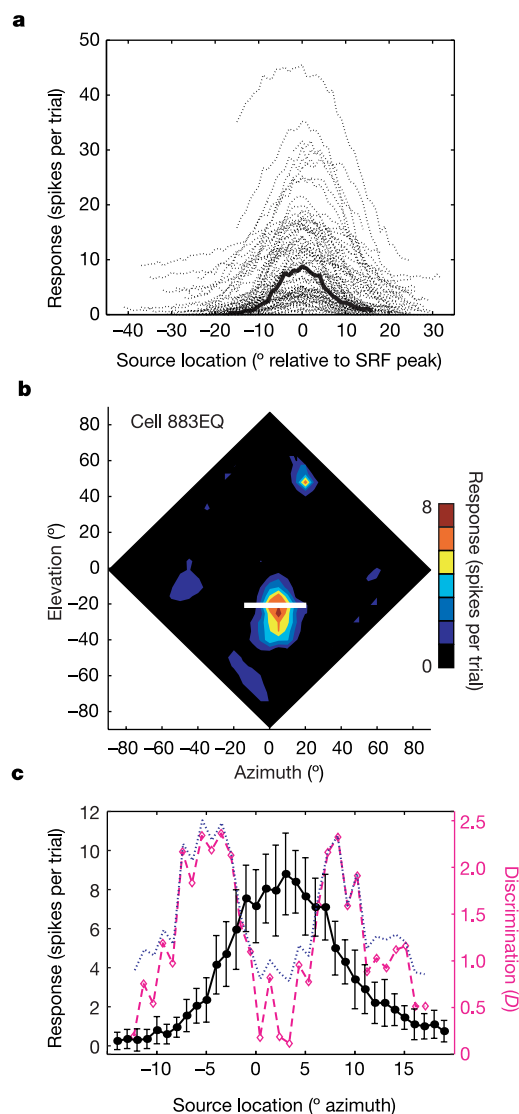


Figure 2 Neuronal discrimination. **a**, Azimuth tuning functions of 83 neurons, aligned such that the peaks are located at 0° azimuth. Unit 883EQ is indicated by the bold line. **b**, SRF of a typical space-map neuron from owl 883. White line represents the azimuthal cross-section assayed at a resolution of 1° . **c**, Neuronal responses are shown in black (mean $\pm$ s.d.). Discrimination performance (magenta) was computed for source pairs separated by 3° . The broken blue line depicts the output of a discrimination model (see text).

locations separated by Δx° , and σ_x and $\sigma_{x+\Delta x}$ refer to the respective s.d. Because sources at x and $x + \Delta x$ will affect neurons differently depending on the locations of their SRF, this method estimates the change in activity across a population of neurons.

Magenta symbols in Fig. 2c represent the D_{neuron} values for all pairs of sources separated by $\Delta x = 3^\circ$. Note that discrimination performance, differs across the extent of the SRF, because it is related to both the change in response and the variance across a 3° separation. Figure 2c shows that this cell could signal 3° changes in source azimuth at suprathreshold levels ($D_{\text{neuron}} \geq 0.8$) at many locations along its SRF. At best, this cell discriminated a 3° separation at a performance of $D_{\text{neuron}} = 2.36$. More than 90% of the units sampled were similar to this example, in that the maximum discrimination performance at a given value of Δx exceeded the owl's behavioural performance. This indicates that the owl does not need to pool responses across coarsely tuned neurons to improve neuronal discrimination^{7–10}. The finding that maximal neuronal acuity exceeds behavioural acuity is inconsistent with the hypothesis that behavioural thresholds are determined by thresholds of the most sensitive neurons^{11,12} and is similar to the findings of studies in cerebral cortex^{13–17}.

As the source separation is increased, neuronal discrimination performance improves. We plotted D_{neuron} values, derived from all 83 neurons, as a function of azimuthal separation (Fig. 3, magenta diamonds). The mean of all of the D_{neuron} values (magenta line), which represents the average change in space-map activity evoked by source displacement, is statistically indistinguishable from the mean behavioural performance (Figs 3 and 4, broken black line; Mann–Whitney U -test; $P > 0.05$). This finding suggests that recovery from habituation or, equivalently, behavioural discrimination performance, is mediated by the change of population activity in the space map.

How could the change in population activity drive the PDR? Although it is premature to propose specific circuitry, we propose a model with the following properties. First, SSNs project to a layer of neurons that habituate to repeated presentation of a stimulus¹⁸, and habituation occurs independently in each neuron^{19,20}. Because firing rates vary within as well as across neurons (Fig. 2a, b), the degree of depression for each input must depend on the accumulated firing rate and variance of the afferent SSN in response to repeated presentation of the habituating stimulus. This is accomplished by normalizing the output of each SSN by the s.d. of these responses.

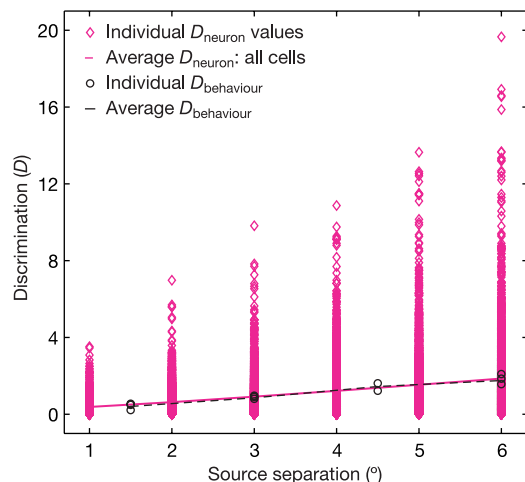


Figure 3 Average neuronal discrimination approximates behavioural performance. At all separations tested, individual neuronal discrimination values (magenta diamonds) can equal or exceed behavioural performance (black circles and line).

Although the response to the habituating stimuli (R_{HS}) declines, the response (R) to stimuli presented from other locations continues (equation (2), Methods).

Second, outputs from the habituating neurons project to an integrator that ultimately controls the magnitude of the PDR. The integrated output is graded, as is the PDR itself, and is directly proportional to behavioural performance. On averaging or scaling by the number of participating neurons (ref. 21 and equation (3), Methods), the integrator output matches behavioural performance (Fig. 4, blue line). Conceptually, this model resembles the proposed novelty detectors in the amygdala²² and would constitute a part of the hippocampal component of the complementary learning systems model²³. The performance of this model, if the space map were composed solely of neurons identical to unit 883EQ, approximates its D_{neuron} values (Fig. 2c, broken blue line). Figure 4 shows the results generated with inputs from all 83 neurons (broken blue line), which yields a good match with behaviour.

Discrimination values derived from single neurons (Fig. 3, magenta diamonds) can be combined in many other ways^{16,24–26}, all of which improve performance. Because the discrimination performance of single neurons can exceed behavioural performance, such pooling schemes yield values that far exceed behavioural performance: optimal pooling²⁴ of individual D_{neuron} values yields $D = 69.23$ for a 3° separation, whereas $D_{\text{behaviour}} = 0.89$. Similarly, discrimination values derived from summing spike rates across all neurons in our sample, which reduces variance, vastly exceed behaviour. Thus, the amount of information present in space-map neurons far exceeds that used for this behaviour, and the owl combines these D_{neuron} values in a non-optimal manner²⁵.

If behaviour could be performed better by single neurons, why would the owl bother with a population representation? Representation of sensory stimuli in populations has inherent advantages, such as elimination of dimensional ambiguity. First, the firing rates of single space-map neurons change in response to alterations not only in stimulus location but also in sound intensity, and such changes produce unambiguous alterations in the collective representation on the space map. Second, a distributed representation is insensitive to the loss of a few constituent neurons. Third, it is conceivable that the information in the space map could be combined in more optimal ways for other tasks, such as sound localization in a multisource environment. Last, our analysis assumes that our discrimination data represent several independent observations of the same stimulus change, and that responses across neurons are uncorrelated. Local connections among neurons exist

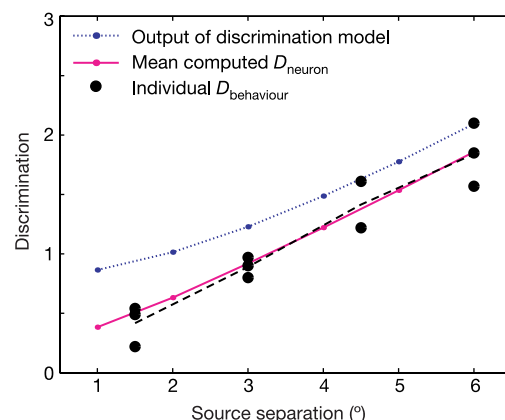


Figure 4 Neuronal computation of spatial discrimination. Discrimination performance by the habituation-based discrimination model (dotted blue line) yields results that approximate behavioural performance.

throughout the auditory system in the owl²⁷, which could lead to some degree of correlation among the responses of space-map neurons. Such correlations would significantly degrade the efficacy of pooling across neurons²⁸.

The change in the neural image, quantified here by measuring changes in the response of space-map neurons caused by azimuthal translocation of a sound source, is an internal 'decision variable' that the PDR circuitry uses to determine the magnitude of a response. It can also function as the decision variable in operant models, such as rating and two-interval forced choice tasks, in which the stimulus at the moment must be compared with a stored representation^{18,22,23,29,30} to generate a response. □

Methods

Owls

Subjects were captive bred adult barn owls of either sex. We used three owls for behavioural assessment (owls 882, 883, 889) and three for neuronal recordings (owls 883, 719, 896). Experimental procedures were approved by the Institutional Animal Care and Use Committee, University of Oregon.

Behaviour

Spatial discrimination behaviour was assessed in a head-fixed preparation by the PDR². The acoustically evoked PDR was habituated by repeated presentation of sounds from a speaker at a single location. Sounds were 100-ms frozen bursts of noise (3–11 kHz) with 5-ms on and off ramps, presented at 52 dBA (A-weighted sound pressure level). This range encompassed frequencies relevant for sound localization in the owl. After habituation (50 sounds, presented once every 12 ± 2 s), sounds were presented from another location that symmetrically straddled the midline. Test stimuli were presented once every 20 trials, resulting in one measurement of the PDR to test stimuli and one measurement of PDR to the habituating stimuli every 4–6 min.

Pupillary dilation was monitored using an infrared pupillometer, which transduced dilation into voltage. The magnitude of the PDR was measured by integrating the output voltage of the pupillometer over 2 s after stimulus onset. The magnitude of the PDR in response to sounds presented from the habituating location—specifically, the magnitude of response of a few randomly selected trials, designated as 'catch trials'—could be then compared with the magnitude of the PDR to sounds from the test location. Typically, 5–7 measurements of each spatial separation were made in a single session lasting 50–70 min. Results were pooled across 4–5 sessions (20–25 observations per separation) to derive discrimination values (see text).

Neurophysiology

Responses of isolated space-specific neurons were recorded in anaesthetized owls. Units could be localized to the exterior nucleus of the inferior colliculus (ICx) on the basis of their response characteristics and location relative to the central nucleus of the inferior colliculus (IC core). Anaesthesia was induced with ketamine and valium, and maintained for the course of the experiment (10–12 h) with nitrous oxide supplemented by isoflurane as needed. We determined each neuron's SRF at a resolution of 10° by presenting broadband bursts (2–11 kHz) from 292 virtual locations.

VAS stimuli were generated by using location-specific filters derived from the head-related transfer functions (HRTFs) of each owl. HRTFs were measured⁶ at a resolution of 5° across the frontal hemisphere (619 locations) and at a resolution of 1° in selected regions of frontal space (245 locations). The sound pressure of VAS stimuli was 50 dBA, which is very close to the level at which behaviour was measured (52 dBA), and was typically 25–30 dB above neuronal threshold. The firing rate of most of these neurons at their best location saturated at 25 dB above threshold. Azimuthal receptive fields were assessed at a resolution of 1° at one of five different elevations: 0°, ±10° and ±20°. Note that the SRFs of space-specific neurons are typically elongated in elevation, and the peak responding area almost always intersected one of the five elevations at which azimuthal discrimination was measured at higher resolution.

Spike counts were measured in a window corresponding to the duration of the sound, delayed by the neuronal response latency. Sounds were presented 20 times from each location, yielding a sample size large enough to assess response variance. For neurons in which we were unable to assess the whole width of the SRF (see the highest responding neuron, Fig. 2c), discrimination was assessed only on the side for which the whole slope, from foot to peak, was sampled. The partially sampled half was not included for analysis.

Neuronal computational model

The output of each space-map neuron is normalized by passage through a linear transformation function, causing the spike rate to be scaled down. This normalized output is then projected to a layer of habituating neurons, where the responses to sound from the habituating stimuli (R_{HS}) decline on repeated presentation. This results in a habituation gradient, where the output of the habituating cell declines to presentation of sounds from the habituating location, but declines less as the test source moves further from the habituating location:

$$\text{Output}_{\text{habituated}} = R - \text{mean}(R_{HS}) \quad (2)$$

The habituated output yields two sets of values, for test and catch trials, from which discrimination can be computed as in equation (1). Note that when responses are

normalized to the s.d. of R_{HS} , the habituated output equals the z-score, which is similar to the computed D (Fig. 3, blue line). The output of each neuron in the habituating layer is summed and averaged²¹:

$$D_{\text{pop}} = (\Sigma_i^n D_{\text{neuron}})/n \quad (3)$$

The resultant output yields a close match with behaviour (Fig. 4, blue line).

Received 27 December 2002; accepted 13 June 2003; doi:10.1038/nature01835.

- Knudsen, E. L., Blasdel, G. G. & Konishi, M. Mechanisms of sound localization in the barn owl (*Tyto alba*). *J. Comp. Physiol. A* **133**, 13–21 (1979).
- Bala, A. D. & Takahashi, T. T. Pupillary dilation response as an indicator of auditory discrimination in the barn owl. *J. Comp. Physiol. A* **186**, 425–434 (2000).
- Knudsen, E. L., Konishi, M. & Pettigrew, J. D. Receptive fields of auditory neurons in the owl. *Science* **198**, 1278–1280 (1977).
- Sakitt, B. Indices of discriminability. *Nature* **241**, 133–134 (1973).
- Sokolov, E. N. Higher nervous functions: the orienting reflex. *Annu. Rev. Physiol.* **25**, 545–580 (1963).
- Keller, C. H., Hartung, K. & Takahashi, T. T. Head-related transfer functions of the barn owl: measurement and neural responses. *Hear. Res.* **118**, 13–34 (1998).
- Hall, J. L. Binaural interaction in the accessory superior-olivary nucleus of the cat. *J. Acoust. Soc. Am.* **37**, 813–823 (1965).
- Baldi, P. & Heiligenberg, W. How sensory maps could enhance resolution through ordered arrangements of broadly tuned receivers. *Biol. Cybern.* **59**, 313–318 (1988).
- Eurich, C. & Swegler, H. Coarse coding: calculation of the resolution achieved by a population of large receptive field neurons. *Biol. Cybern.* **76**, 357–363 (1997).
- Fitzpatrick, D. C., Batra, R., Stanford, T. R., Kuwada, S. A neuronal population code for sound localization. *Nature* **388**, 871–874 (1997).
- Barlow, H. B. Single units and sensation: a neuron doctrine for perceptual psychology? *Perception* **1**, 371–394 (1972).
- Shackleton, T. M., Skottun, B. C., Arnott, R. H. & Palmer, A. R. Interaural time difference discrimination thresholds for single neurons in the inferior colliculus of guinea pigs. *J. Neurosci.* **23**, 716–724 (2003).
- Georgopoulos, A. P., Schwartz, A. B. & Kettner, R. E. Neuronal population coding of movement direction. *Science* **233**, 1416–1419 (1986).
- Newsome, W. T., Britten, K. H. & Movshon, J. A. Neuronal correlates of a perceptual decision. *Nature* **341**, 52–54 (1989).
- Britten, K. H., Shadlen, M. N., Newsome, W. T. & Movshon, J. A. The analysis of visual motion: a comparison of neuronal and psychophysical performance. *J. Neurosci.* **12**, 4745–4765 (1992).
- Shadlen, M. N., Britten, K. H., Newsome, W. T. & Movshon, J. A. A computational analysis of the relationship between neuronal and behavioral responses to visual motion. *J. Neurosci.* **16**, 1486–1510 (1996).
- Recanzone, G. H., Guard, D. C., Phan, M. L. & Su, T. I. Correlation between the activity of single auditory cortical neurons and sound-localization behavior in the macaque monkey. *J. Neurophysiol.* **83**, 2723–2739 (2000).
- Rolls, E. T., Sienkiewicz, Z. J. & Yaxley, S. Hunger modulates the responses to gustatory stimuli of single neurons in the caudolateral orbitofrontal cortex of the macaque monkey. *Eur. J. Neurosci.* **1**, 53–60 (1989).
- Farel, P. B. & Thompson, R. F. Habituation of a monosynaptic response in frog spinal cord: evidence for a presynaptic mechanism. *J. Neurophysiol.* **39**, 661–666 (1976).
- Kandel, E. R., Brunelli, M., Byrne, J. & Castellucci, V. A common presynaptic locus for the synaptic changes underlying short-term habituation and sensitization of the gill-withdrawal reflex in aplysia. *Cold Spring Harb. Symp. Quant. Biol.* **40**, 465–482 (1976).
- Britten, K. H. & Heuer, H. W. Spatial summation in the receptive fields of MT neurons. *J. Neurosci.* **19**, 5074–5084 (1999).
- Rolls, E. T. & Treves, A. *Neural Networks and Brain Function* (Oxford Univ. Press, Oxford, 1998).
- O'Reilly, R. C. & Norman, K. A. Hippocampal and neocortical contributions to memory: advances in the complementary learning systems framework. *Trends Cogn. Sci.* **6**, 505–510 (2002).
- Siebert, W. M. in *Recognizing Patterns* (eds Kolars, P. A. & Eden, M.) 104–133 (MIT Press, Cambridge, MA, 1968).
- Colburn, H. S. Theory of binaural interaction based on auditory-nerve data. I. General strategy and preliminary results on interaural discrimination. *J. Acoust. Soc. Am.* **54**, 1458–1470 (1973).
- Ahissar, M., Ahissar, E., Bergman, H. & Vaadia, E. Encoding of sound-source location and movement: activity of single neurons and interactions between adjacent neurons in the monkey auditory cortex. *J. Neurophysiol.* **67**, 203–215 (1992).
- Fujita, I. & Konishi, M. The role of GABAergic inhibition in processing of interaural time difference in the owl's auditory system. *J. Neurosci.* **11**, 722–739 (1991).
- Zohary, E., Shadlen, M. N. & Newsome, W. T. Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature* **370**, 140–143 (1994).
- Bell, C. C. Memory-based expectations in electrosensory systems. *Curr. Opin. Neurobiol.* **11**, 481–487 (2001).
- Gold, J. I. & Shadlen, M. N. Representation of a perceptual decision in developing oculomotor commands. *Nature* **404**, 390–394 (2000).

Acknowledgements We thank K. Keller for technical assistance, and J. Bala, S. Colburn, P. Dassonville and R. Marrocco for comments on the manuscript. This work was supported by grants from the National Institute on Deafness and Other Communication Disorders and the McKnight Foundation to T.T.T., and from the NIH to M.W.S.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.B. (avinash@uoneuro.uoregon.edu).

Direct modulation of synaptic vesicle priming by GABA_B receptor activation at a glutamatergic synapse

Takeshi Sakaba & Erwin Neher

Department of Membrane Biophysics, Max Planck Institute for Biophysical Chemistry, Göttingen 37077, Germany

Second messenger cascades involving G proteins^{1,2} and calcium³ are known to modulate neurotransmitter release^{4,5}. A prominent effect of such a cascade is the downmodulation of presynaptic calcium influx^{6,7}, which markedly reduces evoked neurotransmitter release^{5,7,8}. Here we show that G-protein-mediated signalling, such as through GABA (γ -amino butyric acid) subtype B (GABA_B) receptors, retards the recruitment of synaptic vesicles during sustained activity and after short-term depression. This retardation occurs through a lowering of cyclic AMP, which blocks the stimulatory effect of increased calcium concentration on vesicle recruitment. In this signalling pathway, cAMP (functioning through the cAMP-dependent guanine nucleotide exchange factor) and calcium/calmodulin cooperate to enhance vesicle priming. The differential modulation of the two forms of synaptic plasticity, presynaptic inhibition and calcium-dependent recovery from synaptic depression, is expected to have interesting consequences for the dynamic behaviour of neural networks.

We characterized neurotransmitter release by using a large glutamatergic synaptic terminal, the calyx of Held^{9,10}, to investigate G-protein-mediated effects other than the well-known reduction of Ca²⁺ currents (I_{Ca}). Figure 1a shows the known effect of a GABA_B receptor agonist, baclofen (20 μ M), on presynaptic I_{Ca} and on the excitatory postsynaptic current (EPSC)^{8,11,12}. In this experiment, both pre- and postsynaptic parts of the calyx of Held synapse were voltage-clamped. A 1-ms depolarizing pulse to +40 mV evoked an average EPSC of $3,066 \pm 734$ pA, and baclofen reduced I_{Ca} to less than 70% (on average to $50.5 \pm 1.7\%$, $n = 9$). Correspondingly, the postsynaptic EPSC was reduced to $9.6 \pm 2.3\%$ of its initial value, in agreement with results at other synapses⁵. Figure 1b shows the time course of the effect of baclofen on both I_{Ca} and the EPSC. The reduction is attributed to a direct action of G proteins on Ca²⁺ channels, because it can be influenced by non-hydrolysable GTP analogues⁸ and G-protein $\beta\gamma$ subunits⁶.

In addition to modulating I_{Ca} , non-hydrolysable GTP analogues also influence the recovery of EPSCs after synaptic depression^{13,14}. We therefore tested whether baclofen influenced recovery from depression after a 300-Hz stimulus train. A pair of 300-Hz trains of action-potential-like stimuli (30 pulses) was applied with an interval of 200 ms (hereafter referred to as the 'control'; Fig. 1c, grey traces), and the same pair of trains was applied after the addition of baclofen (Fig. 1c, black traces). Figure 1c shows that during the second train a fraction of the phasic response (Fig. 1c, grey trace, first ten responses) has recovered, and this recovered fraction is smaller after the addition of baclofen (Fig. 1c, black trace). We assessed the action of baclofen by measuring the integrals of both I_{Ca} and the EPSC. In nine experiments, baclofen reduced the total sum of presynaptic Ca²⁺ influx to $71 \pm 2\%$. The recovered fraction of postsynaptic responses during the second train was $41 \pm 5\%$ in the control and $34.5 \pm 5\%$ under baclofen ($P < 0.05$).

The reduced recovery of the EPSCs under baclofen might be a secondary effect caused by the reduced Ca²⁺ influx, because recovery from synaptic depression is dependent on Ca²⁺ (refs 15–17). In a second series of experiments, we therefore prolonged the individual pulses in the stimulus trains under baclofen to match the

total Ca²⁺ influx between baclofen and the control. Increasing individual pulse lengths under baclofen from 1 ms to 1.5 ms matched this influx exactly (integral of Ca²⁺ influx in trains under baclofen was $100 \pm 7\%$ of the control value in eight such experiments). But the recovered fraction under baclofen was again smaller ($29 \pm 6\%$) than in the control ($40 \pm 6\%$, $P < 0.02$; Fig. 1d). We conclude that vesicle recruitment itself is directly modulated by baclofen. The recovery remained depressed when postsynaptic receptor desensitization and saturation were blocked by cyclothiazide (CTZ) and kynurenic acid (Kyn, $n = 5$).

The metabotropic glutamate receptor (mGluR) agonist LAP4 (50 μ M) and adenosine (50 μ M) also inhibited Ca²⁺ influx during trains to $86 \pm 2\%$ ($n = 6$) and $89 \pm 2\%$ ($n = 3$), respectively¹¹; however, the recovered fraction of postsynaptic responses remained unchanged ($44 \pm 7\%$ versus $43 \pm 7\%$ and $53 \pm 10\%$ versus $56 \pm 10\%$, respectively). The relatively weak effect of these agonists on Ca²⁺ influx indicates either that mGluRs and adenosine receptors may be less effective in stimulating G_o than are GABA_B receptors or that they are tonically activated by endogenous transmitters.

Given the effects of baclofen on recovery from depression, we attempted to map the signalling cascade that is thought to be associated with GABA_B activity. In particular, we looked for interactions between Ca²⁺ and cAMP-mediated effects and for a direct action of baclofen on vesicle recruitment. We applied a stimulation protocol that allows the quantitative study of the

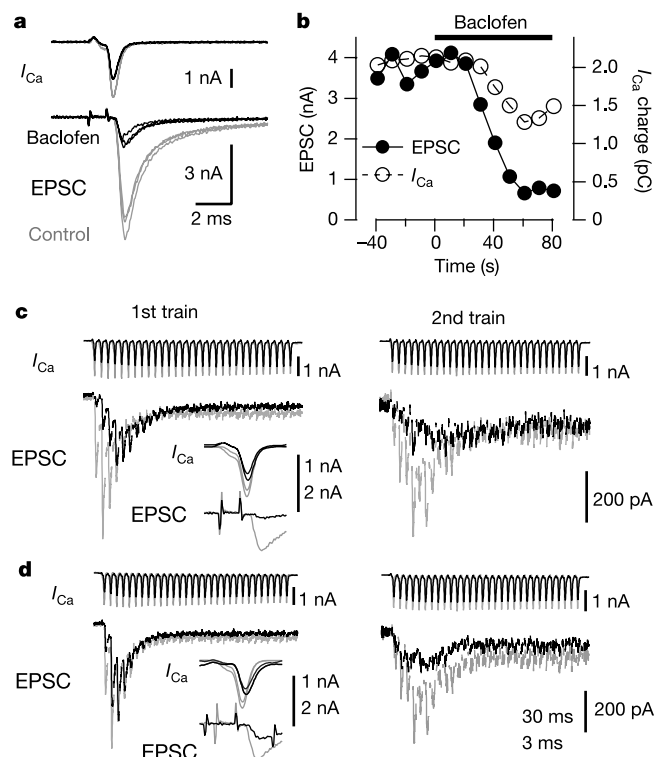


Figure 1 Effects of baclofen on synaptic transmission at the calyx of Held. **a**, A depolarizing pulse (+40 mV for 1 ms) was applied to the presynaptic terminal. Grey traces were obtained during a control period, and black traces were obtained in the presence of baclofen (20 μ M). **b**, Time course of baclofen effect. The cell pair is the same as in **a**. Baclofen was applied at time zero. **c**, A train of depolarizing pulses (30 at 300 Hz) was applied twice with an interval of 200 ms (control, grey traces; baclofen, black traces). Inset shows I_{Ca} in response to the first and the last depolarization during a train, and the EPSCs during the first pulse. **d**, As **c**, but the pulse duration was prolonged from 1 ms to 1.5 ms under baclofen. When two numbers are shown for a scale bar, the bottom number refers to an inset.

Ca^{2+} -dependent recruitment of vesicles¹⁸. In this protocol, a 50-ms depolarization to 0 mV is preceded by a short (2–4 ms) depolarization to +70 mV. The episode at 0 mV has been shown to produce a nearly square pulse of I_{Ca} of about 1.5 nA ($1,733 \pm 33$ pA) and to deplete the pool of releasable vesicles. CTZ (100 μM) and Kyn (1 mM) are added to block desensitization and saturation of the postsynaptic receptors, and the deconvolution method is applied to estimate the cumulative time course of release¹⁸.

Using this protocol, we previously showed that 0.5 mM EGTA in the presynaptic pipette permits the separation of two components of release¹⁸: an initial rapid component with an exponential time course (time constant $\tau_1 = 2.31 \pm 0.12$ ms, $46 \pm 1.6\%$ of the total release), followed by a slow component (time constant $\tau_2 = 29.0 \pm 1.4$ ms). The two components have been interpreted to represent two distinct pools of vesicles that are depleted by the depleting stimulus at timescales given by the two time constants.

Figure 2a shows an experiment using the above stimulation protocol in which the traces from a pair of 50-ms depolarizations have been overlaid (broken lines show traces of the first stimulus in a pair, unbroken lines show those of the second stimulus; the interstimulus interval (ISI) was 1 s). I_{Ca} is indistinguishable between the first and the second depolarization, and the postsynaptic current (EPSC) has recovered to more than 80%. The time course of cumulative release was fitted by a double exponential with time constants similar to those of the first stimulus (the isolated first components are shown as grey traces in Fig. 2a). We repeated such pairs of depolarizations while varying the ISI, and plotted the recovery time course of the first component as the recovered fraction (mean values from nine experiments) against the ISI (Fig. 2d). A double exponential with time constants of 292 ms (71%) and 4.7 s (29%, broken trace in Fig. 2d) was fitted to the recovery time course. A similar plot for the second component showed that it recovered very quickly (Fig. 2e). This component did not change in any great way during any of the manipulations carried out in this study.

We used this protocol to investigate further the modulatory effects of baclofen. Baclofen was applied immediately after the whole-cell configuration was established (Fig. 2b). Presynaptic I_{Ca} amplitudes ($1,368 \pm 50$ pA) were within the range of control values, because prepulses (+70 mV for 4 ms) effectively relieved G-protein-mediated inhibition of I_{Ca} (ref. 6). Baclofen slowed the recovery of the first component (half-recovery at 1.53 ± 0.30 s ISI versus 0.42 ± 0.08 s ISI in the control, $P < 0.01$, Fig. 2d). Recovery of the second component was not modulated by baclofen (Fig. 2e).

As the effects of GABA_B-type agonists can be caused by inhibition of adenylate cyclase, we examined whether exogenously applied cAMP would reverse the effect of baclofen on release. Addition of cAMP to the patch pipette prevented the inhibitory effect of baclofen on vesicle recruitment (Fig. 2c, d). In seven terminals, half-recovery occurred at 0.66 ± 0.14 s. In one terminal, the recovery time course remained similar to that under baclofen. Thus, baclofen, possibly through G_o (ref. 6), seems to decrease cAMP concentrations in the terminal, thereby slowing down vesicle recruitment.

The data presented so far could be explained most readily if it is assumed that there is tonic activation of adenylate cyclase under control conditions. We therefore carried out experiments in which we tried to inhibit adenylate cyclase artificially. We used the non-hydrolysable guanosine nucleotides GDP- β S and GTP- γ S, which lock heterotrimeric G proteins in inactive and active states, respectively. Because the basal activity of inhibitory G proteins is low (as assayed by tonic inhibition of I_{Ca} ; ref. 19), the effect of GDP- β S should be primarily a block of either G_s or small G proteins.

We found that the recruitment of vesicles to the rapidly releasing pool was strongly retarded when we included 3 mM GDP- β S in the presynaptic patch pipette, in agreement with the expectation that cAMP should be lowered. This block was even stronger than that

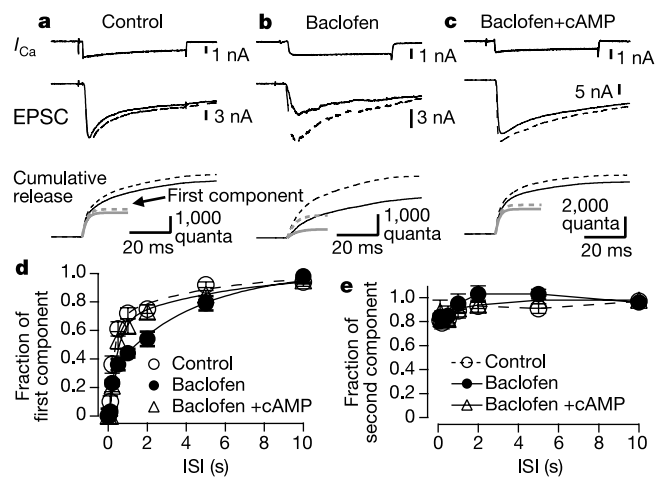


Figure 2 Inhibitory effect of baclofen on vesicle recruitment. **a**, The presynaptic terminal was depolarized to +70 mV for 2 ms and then repolarized to 0 mV for 50 ms (broken traces). The same pulse was applied a second time (unbroken traces) at an ISI of 1 s. I_{Ca} , EPSC and the cumulative release are shown. Grey traces indicate the fitted time courses of the first component. **b**, As **a**, but baclofen (20 μM) was applied immediately after the whole-cell recording started. **c**, As **b**, but 1 mM cAMP was added to the presynaptic pipette solution. **d**, Recovered fraction of the first component plotted against ISI. Recovery time constants of double exponential fits are 292 ms (71%), 4.7 s (control); 239 ms (32%), 4.3 s (baclofen); and 421 ms (71%), 5.8 s (baclofen plus cAMP). **e**, Recovered fraction of the second component plotted against ISI.

under baclofen. Figure 3a shows an extreme example of this, in which the EPSC during a second depolarization at 1-s ISI is strongly reduced (unbroken traces). The recovery time course of the rapidly releasing component (Fig. 3b, squares; $n = 6$) shows the blocking effect of GDP- β S clearly for ISIs between 200 ms and 5 s (half-recovery at 2.47 ± 0.64 s). We did not observe any consistent acceleration of recruitment when we used GTP- γ S (0.5 mM) in the pipette (recovery time constants of 410 ms, 69%; and 4.5 s, 31%; $n = 7$). This again is compatible with the assumption that a G-protein-mediated signalling pathway that maintains cAMP and supports rapid vesicle recruitment is tonically activated²⁰.

If basal G-protein activity stimulates adenylate cyclase, exogenously applied cAMP, which is expected to be reduced under GDP- β S, will reverse the effect. In contrast to GDP- β S alone, when 3 mM GDP- β S and 1 mM cAMP were applied together, the postsynaptic response recovered normally (Fig. 3c). A plot of the recovered fraction of the first component as a function of ISI showed that its recovery under GDP- β S plus cAMP was about as fast as that of the control (half-recovery at 0.69 ± 0.07 s, $n = 8$; Fig. 3b). cAMP alone had a slight depressing effect, if any, on the recovery (Fig. 3b and Supplementary Information).

To substantiate further the role of adenylate cyclase in priming, we carried out experiments with an adenylate cyclase inhibitor MDL 12330A (1 mM) and found that the data points nearly superimposed on the recovery time course obtained in experiments with GDP- β S (Fig. 3b, filled triangles). The effect of MDL 12330A could be reversed by 1 mM cAMP ($n = 3$). These data complement our previous findings showing that stimulation of adenylate cyclase leads to a slight increase in the fast component of release²¹. The general kinase inhibitor staurosporine (2 μM ; ref. 22) and protein kinase A inhibitor 6-22 amide (50 μM) did not block recovery of the first component, suggesting that protein kinase A (PKA) is not involved in vesicle priming.

As most regulatory effects of cAMP are mediated by PKA, we considered what other targets of cAMP might be involved in its action on vesicle recruitment. We loaded the presynaptic terminal with 8CPT-2Me-cAMP (100 μM), a specific activator of cAMP-

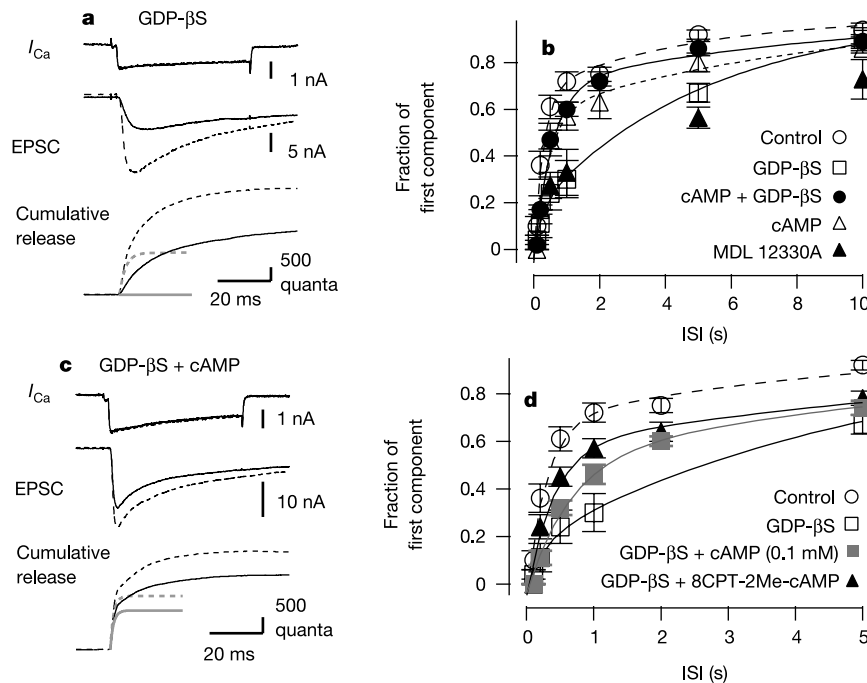


Figure 3 Supplementation of cAMP prevents the inhibitory effect of GDP- β S. The protocol was the same as in Fig. 2a, but the presynaptic patch pipette contained 3 mM GDP- β S (**a**) or 3 mM GDP- β S plus 1 mM cAMP (**c**), instead of 0.3 mM GTP. **b**, Recovered amount of the first component plotted against the ISI. Recovery time constants are 273 ms (18%),

5.2 s (3 mM GDP- β S); 545 ms (74%), 8.3 s (3 mM GDP- β S plus 1 mM cAMP); and 429 ms (64%), 8.3 s (1 mM cAMP). **d**, Recovered amount of the first component plotted against ISI. Recovery time constants are 400 ms (62%), 8.3 s (3 mM GDP- β S plus 0.1 mM 8CPT-2Me-cAMP); and 645 ms (56%), 8.3 s (3 mM GDP- β S plus 0.1 mM cAMP).

dependent guanosine exchange factor (cAMP-GEF; also known as EPAC2)²³, together with 3 mM GDP- β S (Supplementary Information). 8CPT-2Me-cAMP effectively reversed the effect of GDP- β S (half-recovery at 0.86 ± 0.15 s, $n = 7$ cell pairs; Fig. 3d, filled triangles). The same concentration of cAMP was less effective in reversing the effect of GDP- β S (half-recovery at 1.35 ± 0.25 s, $n = 6$; Fig. 3d, grey squares). If the target of the adenine nucleotides were PKA, one would expect cAMP to be many times more effective than 8CPT-2Me-cAMP²³. Thus, one of the isoforms of cAMP-GEF is likely to be involved in vesicle priming.

It has been shown previously that the rapid recovery of release after long depolarizing pulses—similar to the recovery after trains of afferent nerve stimulation—is dependent on the increased concentration of Ca^{2+} that prevails after strong stimulation^{15–18}. We therefore considered whether cAMP could overcome this Ca^{2+} requirement and enhance vesicle recruitment by itself or whether the two stimuli might depend on each other. Addition of 5 mM EGTA by itself retarded recovery after prolonged depolarization (see Supplementary Information). When 1 mM cAMP plus 5 mM EGTA were added to the patch pipette, only a marginal acceleration of recovery was observed. Thus, cAMP is ineffective when a low concentration of Ca^{2+} is maintained by EGTA. We conclude that Ca^{2+} /calmodulin¹⁸ does not exert its effect on vesicle recruitment by stimulating adenylate cyclase. Rather, cAMP and Ca^{2+} are both required at a late step in vesicle recruitment.

Our results show that baclofen has a dual mechanism of action in modulating transmitter release through G proteins. One effect, which has been well-characterized^{15,7,8}, results in a reduction of Ca^{2+} influx. The other results in a slowdown of vesicle recruitment during and after repetitive activity. Whereas the effect on Ca^{2+} influx is mediated directly by G-protein $\beta\gamma$ subunits⁶, the inhibitory action on vesicle recruitment is exclusively mediated by a decrease in cAMP concentration. However, cAMP seems to act downstream of heterotrimeric G proteins, in concert with Ca^{2+} at a vesicle priming step. Its action is probably mediated through cAMP-GEFII, which

interacts with Rim²⁴. The involvement of Rim, as a target of cAMP, would explain why Ca^{2+} /calmodulin and cAMP have to cooperate to exert their effect. Rim interacts with Munc13 (ref. 25), which has a calmodulin interaction site²⁶, and both Rim and Munc13 have been associated with vesicle priming in genetic and molecular studies^{24,25,27}.

In view of the many actions of second messengers that have been proposed to be associated with G-protein-mediated modulation of neurotransmitter release, we note that the modulation is specific in a triple sense. First, prominent downmodulation of I_{Ca} is most important for all forms of synchronous release, during both single spikes and spike trains, and it is mediated mainly through G-protein $\beta\gamma$ subunits. Second, the newly described modulatory effect of baclofen on vesicle recruitment of the fast vesicle pool is physiologically relevant for synchronous activity during trains of action potentials. Last, the slow component of release contributes strongly to asynchronous release during repetitive activity (E.N. *et al.*, unpublished) and remains unchanged by all of the modulatory mechanisms. In regard to spontaneous release, which we did not address here, it is well established that G-protein signalling pathways modulate this process directly^{28–30}. It will be interesting to explore how this rich repertoire of differential regulation is used to tune the dynamic properties of neuronal networks. □

Methods

Brain samples

Brainstem slices (150–200 μ m thick) were prepared as described^{9,10}. In short, brainstem slices from Wistar rats aged 8–11 d were cut in ice-cold, low-calcium saline containing 125 mM NaCl, 2.5 mM KCl, 0.1 mM $CaCl_2$, 3 mM $MgCl_2$, 25 mM glucose, 25 mM $NaHCO_3$, 1.25 mM NaH_2PO_4 , 0.4 mM ascorbic acid, 3 mM myo-inositol and 2 mM sodium pyruvate (pH 7.3–7.4, 320 mOsm) and bubbled with 95% O_2 and 5% CO_2 . Slices were incubated in the chamber for at least 30 min at 36 °C in normal extracellular saline. Normal extracellular saline was the same as low-calcium saline, except that 2.0 mM $CaCl_2$ and 1.0 mM $MgCl_2$ were included. We carried out experiments within 4 h of preparing the slices. Experiments were performed according to the ethical guidelines of the state of Lower Saxony.

Recordings

All recordings were done at room temperature (around 21–24 °C). A calyx of Held and the postsynaptic MNTB principal neuron were whole-cell-clamped at –80 mV with an EPC9/2 amplifier (HEKA). In the experiments of Figs 2 and 3, 100 μ M CTZ and 1 mM Kyn were added to block desensitization and saturation of AMPA (α -amino-3-hydroxy-5-methyl-4-isoxazole propionic acid) receptors. The presynaptic patch pipette (3–5 M Ω) contained 135 mM caesium gluconate, 20 mM TEA-Cl, 10 mM HEPES, 5 mM sodium phosphocreatine, 4 mM MgATP, 0.3 mM GTP and 0.05 mM EGTA (pH 7.2). In the experiments of Figs 2 and 3, 0.5 mM EGTA was added to the pipette to separate the first and the second components of release. In some experiments, we replaced 0.3 mM GTP by GTP analogues. The presynaptic series resistance (5–20 M Ω) was compensated by 30–90%. The postsynaptic pipette (2–3.5 M Ω) contained the same solution as the presynaptic pipette, except that EGTA was increased to 5 mM. The postsynaptic series resistance (3–8 M Ω) was compensated by the amplifier so that the uncompensated resistance was below 3 M Ω . The remaining resistance was further compensated off-line. We obtained 8CPT-2Me-cAMP, CTZ and D-AP5 from Tocris, and GTP- γ S, GDP- β S, cAMP, MDL 12330A and PKI (protein kinase A inhibitor 6-22 amido) from Calbiochem.

Release rates and statistical measures

Quantal release rates were estimated by the deconvolution method adapted for the calyx of Held¹⁸. Release rates, as determined by deconvolution, were integrated to obtain cumulative release. All values are expressed as the mean $\pm$ s.e.m. Statistical significance was determined by paired and unpaired *t*-tests. Cumulative release was fitted with double exponentials. In the double-pulse protocol, the second pulse was fitted with a double exponential either by allowing for two time constants as free parameters, or by fixing the time constants to the same values as used in the first pulse. Both methods gave satisfactory fits. Time courses of the amplitude of the two components (recovered fraction) were fitted by double exponentials.

Received 31 March; accepted 13 June 2003; doi:10.1038/nature01859.

- Simon, M. I., Strathmann, M. P. & Gautam, N. Diversity of G proteins in signal transduction. *Science* **252**, 802–808 (1991).
- Gilman, A. G. G proteins and regulation of adenyl cyclase. *Biosci. Rep.* **15**, 65–97 (1995).
- Zucker, R. S. & Regehr, W. G. Short-term synaptic plasticity. *Annu. Rev. Physiol.* **64**, 355–405 (2002).
- Nicoll, R. A., Malenka, R. C. & Kauer, J. A. Functional comparison of neurotransmitter receptor subtypes in mammalian central nervous system. *Physiol. Rev.* **70**, 513–565 (1990).
- Wu, L. G. & Saggau, P. Presynaptic inhibition of elicited neurotransmitter release. *Trends Neurosci.* **20**, 204–212 (1997).
- Kajikawa, Y., Saitoh, N. & Takahashi, T. GTP-binding protein $\beta\gamma$ subunits mediate presynaptic calcium current inhibition by GABA_A receptor. *Proc. Natl Acad. Sci. USA* **98**, 8054–8058 (2001).
- Dittman, J. S. & Regehr, W. G. Contributions of calcium-dependent and calcium-independent mechanisms to presynaptic inhibition at a cerebellar synapse. *J. Neurosci.* **16**, 1623–1633 (1996).
- Takahashi, T., Kajikawa, Y. & Tsujimoto, T. G-protein-coupled modulation of presynaptic calcium currents and transmitter release by a GABA_A receptor. *J. Neurosci.* **18**, 3138–3146 (1998).
- Forsythe, I. D. Direct patch recording from identified presynaptic terminals mediating glutamatergic EPSCs in the rat CNS, *in vitro*. *J. Physiol. (Lond.)* **479**, 381–387 (1994).
- Borst, J. G. G., Helmchen, F. & Sakmann, B. Pre- and postsynaptic whole-cell recordings in the medial nucleus of the trapezoid body of the rat. *J. Physiol. (Lond.)* **489**, 825–840 (1995).
- Barnes-Davies, M. B. & Forsythe, I. D. Pre- and postsynaptic glutamate receptors at a giant synapse in rat auditory brainstem slices. *J. Physiol. (Lond.)* **488**, 387–406 (1995).
- Isaacson, J. S. GABA_A receptor-mediated modulation of presynaptic currents and excitatory transmission at a fast central synapse. *J. Neurophysiol.* **80**, 1571–1576 (1998).
- Takahashi, T., Hori, T., Kajikawa, Y. & Tsujimoto, T. The role of GTP-binding protein activity in fast central synaptic transmission. *Science* **289**, 460–463 (2000).
- Hess, S. D., Doroshenko, P. A. & Augustine, G. J. A functional role for GTP-binding proteins in synaptic vesicle cycling. *Science* **259**, 1169–1172 (1993).
- Dittman, J. S. & Regehr, W. G. Calcium dependence and recovery kinetics of presynaptic depression at the climbing fiber to Purkinje cell synapse. *J. Neurosci.* **18**, 6147–6162 (1998).
- Stevens, C. F. & Wesseling, J. F. Activity-dependent modulation of the rate at which synaptic vesicles become available to undergo exocytosis. *Neuron* **21**, 415–424 (1998).
- Wang, L. Y. & Kaczmarek, L. K. High-frequency firing helps replenish the readily releasable pool of synaptic vesicles. *Nature* **394**, 384–388 (1998).
- Sakaba, T. & Neher, E. Calmodulin mediates rapid recruitment of fast-releasing synaptic vesicles at a calyx-type synapse. *Neuron* **32**, 1119–1131 (2001).
- Cuttler, M. E., Tsujimoto, T., Forsythe, I. D. & Takahashi, T. Facilitation of the presynaptic calcium current at an auditory synapse in rat brainstem. *J. Physiol. (Lond.)* **512**, 723–729 (1998).
- Seifert, R. & Wenzel-Seifert, K. Constitutive activity of G-protein-coupled receptors: cause of disease and common property of wild-type receptors. *Naunyn-Schmiedeberg's Arch. Pharmacol.* **366**, 381–416 (2002).
- Sakaba, T. & Neher, E. Preferential potentiation of fast-releasing synaptic vesicles by cAMP at the calyx of Held. *Proc. Natl Acad. Sci. USA* **98**, 331–336 (2001).
- Sakaba, T. & Neher, E. Involvement of actin polymerization in vesicle recruitment at the calyx of Held synapse. *J. Neurosci.* **23**, 837–846 (2003).
- Enserink, J. M. *et al.* A novel Epac-specific cAMP analogue demonstrates independent regulation of Rap1 and ERK. *Nature Cell Biol.* **4**, 901–906 (2002).
- Ozaki, N. *et al.* cAMP-GEF II is a direct target of cAMP in regulated exocytosis. *Nature Cell Biol.* **2**, 805–811 (2000).
- Betz, A. *et al.* Functional interaction of the active zone protein Munc13-1 and Rim1 in synaptic vesicle priming. *Neuron* **30**, 183–196 (2001).
- Xu, X. Z. S. *et al.* Retinal targets for calmodulin include proteins implicated in synaptic transmission. *J. Biol. Chem.* **273**, 31297–31307 (1998).
- Eliasson, L. *et al.* SUR1 regulates PKA-independent cAMP-induced granule priming in mouse pancreatic B-cells. *J. Gen. Physiol.* **121**, 181–197 (2003).

- Silinsky, E. M. On the mechanism by which adenosine receptor activation inhibits the release of acetylcholine from motor-nerve endings. *J. Physiol. (Lond.)* **346**, 243–256 (1984).
- Dale, N. & Kandel, E. R. Facilitatory and inhibitory transmitters modulate spontaneous transmitter release at cultured Aplysia sensorimotor synapses. *J. Physiol. (Lond.)* **421**, 203–222 (1990).
- Scanziani, M., Capogna, M., Gähwiler, B. H. & Thompson, S. M. Presynaptic inhibition of miniature excitatory synaptic currents by baclofen and adenosine in the hippocampus. *Neuron* **9**, 919–927 (1992).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank N. Brose, F. Felmy, C. Rosenmund, F. Sachs, R. Schneggenburger, G. Schultz and M. Wölfel for critical comments on the manuscript. This work was supported in part by a grant from the Deutsche Forschungsgemeinschaft (DFG Research Center on Molecular Physiology of the Brain).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E.N. (eneher@gwdg.de).

P2X₄ receptors induced in spinal microglia gate tactile allodynia after nerve injury

Makoto Tsuda^{1,2}, Yukari Shigemoto-Mogami¹, Schuichi Koizumi³, Akito Mizokoshi⁴, Shinichi Kohsaka⁵, Michael W. Salter² & Kazuhide Inoue^{1,4}

¹Division of Biosignaling, National Institute of Health Sciences, 1-18-1 Kamiyoga, Setagaya, Tokyo 158-8501, Japan

²Programme in Brain and Behaviour, The Hospital for Sick Children, 555 University Avenue, Toronto, Ontario M5G 1X8, Canada

³Division of Pharmacology, National Institute of Health Sciences, 1-15-1 Kamiyoga, Setagaya, Tokyo 158-8501, Japan

⁴Department of Molecular and System Pharmacology, Graduate School of Pharmaceutical Sciences, Kyushu University, 3-1-1 Maidashi, Higashi, Fukuoka 812-8582, Japan

⁵Department of Neurochemistry, National Institute of Neuroscience, 4-1-1 Ogawa-higashi, Kodaira, Tokyo 187-8502, Japan

Pain after nerve damage is an expression of pathological operation of the nervous system^{1,2}, one hallmark of which is tactile allodynia—pain hypersensitivity evoked by innocuous stimuli. Effective therapy for this pain is lacking, and the underlying mechanisms are poorly understood. Here we report that pharmacological blockade of spinal P2X₄ receptors (P2X₄Rs)^{3–7}, a subtype of ionotropic ATP receptor⁸, reversed tactile allodynia caused by peripheral nerve injury without affecting acute pain behaviours in naive animals. After nerve injury, P2X₄R expression increased strikingly in the ipsilateral spinal cord, and P2X₄Rs were induced in hyperactive microglia but not in neurons or astrocytes. Intraspinal administration of P2X₄R antisense oligodeoxynucleotide decreased the induction of P2X₄Rs and suppressed tactile allodynia after nerve injury. Conversely, intraspinal administration of microglia in which P2X₄Rs had been induced and stimulated, produced tactile allodynia in naive rats. Taken together, our results demonstrate that activation of P2X₄Rs in hyperactive microglia is necessary for tactile allodynia after nerve injury and is sufficient to produce tactile allodynia in normal animals. Thus, blocking P2X₄Rs in microglia might be a new therapeutic strategy for pain induced by nerve injury.

We investigated the mechanisms underlying tactile allodynia using a model involving injury to the fifth lumbar (L5) spinal nerve⁹. Animals with L5 nerve injury but not sham controls (not shown) displayed marked tactile allodynia: the paw withdrawal

threshold decreased from 15.1 ± 0.1 g ($n = 15$ animals) before the injury to 3.1 ± 0.3 g ($n = 15$ animals) at day 7 ($P < 0.001$; Fig. 1a) and 2.4 ± 0.2 g ($n = 15$ animals) at day 14 ($P < 0.001$; Fig. 1c). We tested for the involvement of P2X receptors (P2XRs) in the tactile allodynia by using 2',3'-O-(2,4,6-trinitrophenyl) adenosine 5-triphosphate (TNP-ATP)^{8,10}, an antagonist of P2XR subtypes P2X₁₋₄R. We found that after intrathecal injection of TNP-ATP (30 nmol) the paw withdrawal threshold increased gradually, peaked at about 45 min after the injection and then returned to the pre-injection level over the next 45 min. When 30 nmol TNP-ATP was administered on day 7, the paw withdrawal threshold 45 min after injection was 14.6 ± 0.5 g ($n = 6$ animals; $P < 0.001$). The paw withdrawal threshold at the peak of the effect of TNP-ATP was not different from that before nerve injury; tactile allodynia was therefore reversed by 30 nmol TNP-ATP on day 7. On day 14 the paw withdrawal threshold was 11.6 ± 1.3 g ($n = 6$ animals) 45 min after injection of TNP-ATP ($P < 0.001$), which represented a 73% recovery of paw withdrawal threshold. Intrathecal administration of the vehicle (PBS) had no effect on either testing day (Fig. 1). The increase in paw withdrawal threshold by TNP-ATP was dose-dependent; the dose producing half-maximal effect calculated as 8.1 nmol on day 7 and 15.5 nmol on day 14 (Fig. 1b, d). We observed no alteration in motor behaviour after TNP-ATP administration (data not shown). Nor did TNP-ATP affect the paw withdrawal threshold on the side contralateral to the nerve injury: 45 min after administration of PBS the threshold was 15.1 ± 0.1 g compared with 14.3 ± 0.8 g 45 min after administration of 30 nmol TNP-ATP ($P > 0.1$). These results together indicate that TNP-ATP caused a dose-dependent, reversible recovery of paw withdrawal threshold on the nerve-injured side without a non-specific effect on motor or sensory functioning. Thus, tactile allodynia produced by nerve

injury seems to be mediated at the spinal level by a P2XR sensitive to TNP-ATP.

We next tested pyridoxal-phosphate-6-azophenyl-2',4'-disulphonic acid (PPADS), an antagonist of P2XR subtypes P2X_{1,2,3,5,7}R but not of P2X₄R (ref. 8), which has antinociceptive effects in models of acute and inflammatory pain^{11,12}. We found that intrathecal administration of 30 or 100 nmol PPADS had no effect on paw withdrawal threshold on either day 7 (Fig. 1a, b) or day 14 (Fig. 1c, d). At these intrathecal doses, PPADS is known to suppress nociceptive behaviours caused by intrathecal injection of the P2X_{1,3}R agonist α,β -methylene ATP¹³, and those behaviours during the second phase of the formalin test¹¹ or evoked by the injection of bee venom¹², both of which are models of inflammatory pain. Thus, these are doses of PPADS that should have been sufficient to increase paw withdrawal threshold in the present study if the P2XRs involved in tactile allodynia were PPADS-sensitive. The lack of effect of PPADS on paw withdrawal threshold together

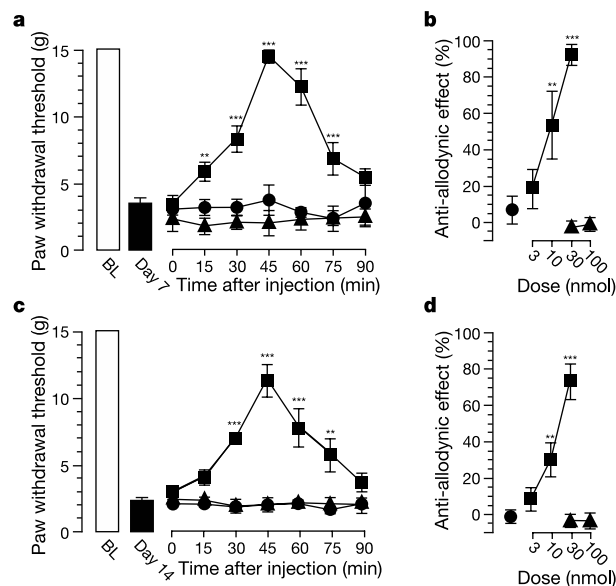


Figure 1 Spinal administration of TNP-ATP but not PPADS reverses tactile allodynia caused by injury to the L5 spinal nerve. The withdrawal threshold of tactile stimulation to the ipsilateral hindpaw was examined by applying von Frey filaments. Symbols in all panels: squares, TNP-ATP; triangles, PPADS; circles, PBS. **a, c**, Paw withdrawal threshold (mean $\pm$ s.e.m.) before nerve injury (BL), 7 days (**a**; day 7) and 14 days (**c**; day 14) after nerve injury (day 7 and day 14, $P < 0.001$ compared with BL). The 30 nmol line graphs show the effects of intrathecal administration of TNP-ATP and PPADS on the decrease in paw withdrawal threshold 7 (**a**) and 14 days (**c**) after nerve injury. **b, d**, Anti-allodynic effect (mean $\pm$ s.e.m.) of TNP-ATP 7 days (**b**) and 14 days (**d**) after nerve injury. Anti-allodynic effect (%) = $100(\text{test value} - \text{pre-injection value})/(\text{value at } 15.1 \text{ g} - \text{pre-injection value})$. Significance compared with the PBS-treated group: two asterisks, $P < 0.01$; three asterisks, $P < 0.001$.

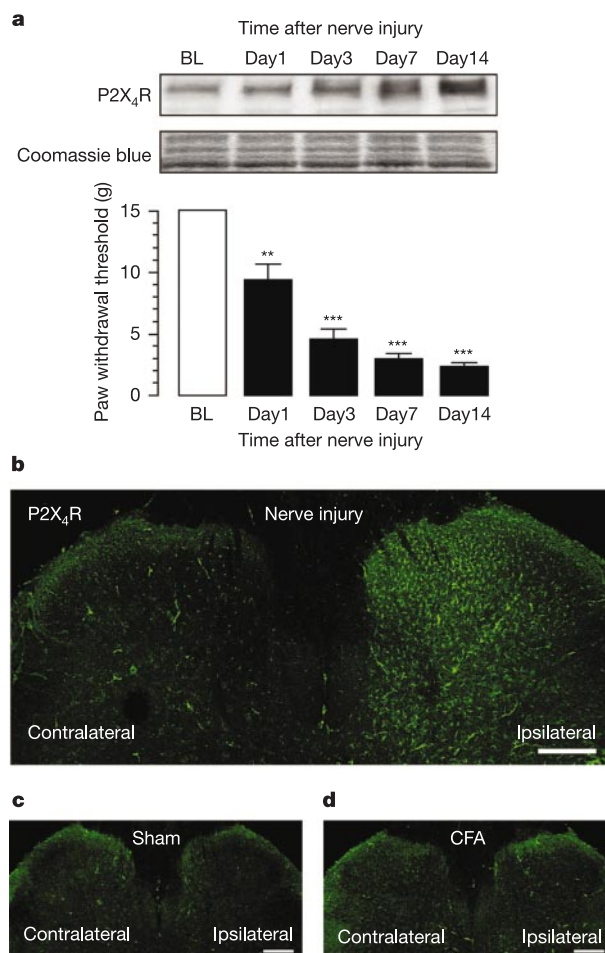


Figure 2 Marked upregulation of P2X₄R level in the spinal dorsal horn after injury to the L5 nerve. **a**, Western blot analysis of P2X₄R protein detected by anti-P2X₄R antibody in the membrane fraction from the spinal cord ipsilateral to the nerve injury at different time (top panel). The total protein loaded on each lane was stained with Coomassie blue (middle panel). The time course of change in P2X₄R protein is similar to that in paw withdrawal threshold (bottom panel). Significance compared with the pre-injury baseline (BL): two asterisks, $P < 0.01$; three asterisks, $P < 0.001$. **b-d**, Visualization of the P2X₄R protein detected by P2X₄R antibody in the L5 dorsal spinal cord by immunofluorescence analysis with confocal microscopy. Photographs show the P2X₄R immunofluorescence in the dorsal horn 14 days after nerve injury (**b**), 14 days after sham operation (**c**) and 7 days after the injection of CFA into the plantar surface of the hindpaw, an inflammatory pain model (**d**). Scale bars, 200 μ m.

with the increase by TNP-ATP indicates that tactile allodynia caused by L5 nerve injury depends upon spinal P2XRs that are sensitive to TNP-ATP and insensitive to PPADS. The pharmacological profile of these P2XRs is consistent with that of the P2X₄R subtype^{3–8}; we therefore further explored the role of P2X₄R in tactile allodynia after nerve injury.

We examined the level of P2X₄R protein in homogenates from spinal cord of naive and nerve-injured rats and found that P2X₄R protein in the ipsilateral spinal cord increased markedly after injury to the L5 nerve (Fig. 2a): the increase in P2X₄R was detected as early as day 1 and the highest level was observed on day 14. In contrast, the level of P2X₄R protein in the contralateral spinal cord was not different on either day 7 or day 14 from that in naive rats (not shown). The time course of the change in P2X₄R level in the spinal cord and the bilateral difference in P2X₄R levels match the emergence of the tactile allodynia (Fig. 2a, bottom panel). To examine the distribution of P2X₄R, we performed immunofluorescence on sections of the L5 spinal dorsal horn (Fig. 2b–d). In the spinal cord ipsilateral to the nerve injury, we observed strong, punctate P2X₄R immunofluorescence dotted throughout the dorsal horn (Fig. 2b); the punctate labelling observed at low magnification was

due to immunofluorescence of individual cells (see below). In contrast, P2X₄R immunofluorescence was weaker and much less extensive in the dorsal horn contralateral to the nerve injury (Fig. 2b) or in that of sham-operated rats (Fig. 2c). The immunofluorescence was abolished by preabsorbing the antibody with the immunogen peptide for P2X₄R antibody, indicating that the observed staining was not a non-specific signal. Thus, the level of P2X₄R increased markedly in the dorsal horn ipsilateral to the nerve injury with a time course matching that of the development of tactile allodynia. In contrast to this increase in P2X₄R level, we found that there was no change in P2X₄R immunofluorescence in the dorsal horn 7 days after intraplantar injection of complete Freund's adjuvant (CFA; Fig. 2d). CFA produces sustained inflammation of the paw and prolonged hypersensitivity to pain. Thus, nerve injury but not persistent peripheral inflammation caused an increase in P2X₄R level in the dorsal horn.

To identify the type of cell expressing P2X₄R after nerve injury, on day 14 we performed double immunofluorescence labelling for P2X₄R and for cell-type-specific markers: for neurons, microtubule-associated protein 2 (MAP2) and neuronal nuclei (NeuN); for astrocytes, glial fibrillary acidic protein (GFAP); and for microglia,

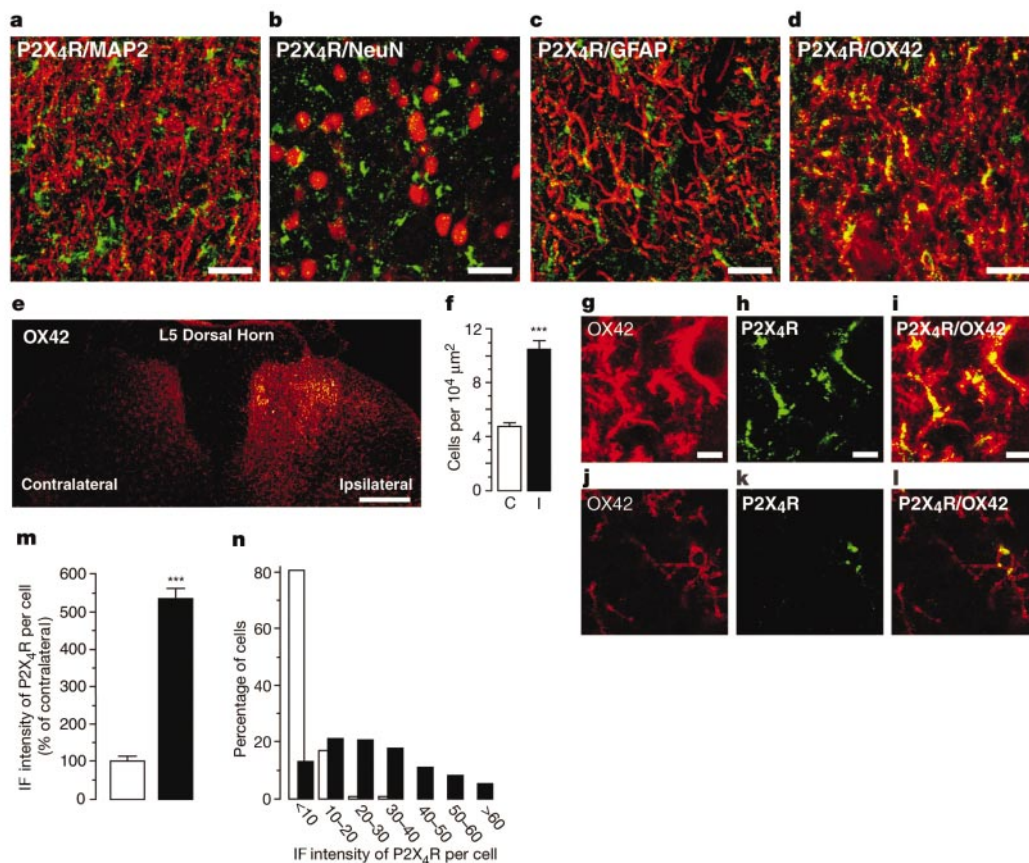


Figure 3 P2X₄R is induced in hyperactive microglia but not in neurons or astrocytes. All experiments were performed with spinal cord sections 14 days after nerve injury. **a–d**, Double immunofluorescence labelling of P2X₄R with MAP2 (**a**) and NeuN (**b**), markers of neurons; GFAP (**c**), a marker of astrocytes; and OX42 (**d**), a marker of microglia. Most P2X₄R-positive cells are double-labelled (yellow) with OX42 (**d**). Scale bars, 50 μm. The remaining small proportion (less than 3%) of P2X₄R-positive cells were double-labelled with ED2, a marker of perivascular macrophages (not shown). **e**, OX42 labelling after nerve injury in the dorsal horn. Scale bar, 200 μm. **f**, The number of cells per 10⁴ μm² labelled with OX42 (mean ± s.e.m.) in the dorsal horn ipsilateral (I) and contralateral (C) to the nerve injury (three asterisks, $P < 0.001$ compared with contralateral). **g–l**, P2X₄R immunofluorescence in hyperactive microglia labelling

intensely with OX42 in the ipsilateral dorsal horn (**g** (OX42, red), **h** (P2X₄R, green), **i** (merged, yellow)) and in resting microglia labelling much less with OX42 in contralateral dorsal horn (**j** (OX42, red), **k** (P2X₄R, green), **l** (merged, yellow)). Scale bars, 10 μm. **m**, Immunofluorescence (IF) intensity of P2X₄R protein in individual microglia determined as the average pixel density in the ipsilateral (filled bar; $n = 236$ OX42-positive cells) and contralateral (open bar; $n = 104$ OX42-positive cells) dorsal horn (three asterisks, $P < 0.001$). Results are percentages (mean ± s.e.m.) of the values normalized to the mean values of the contralateral dorsal horn. **n**, Histogram of the percentage of dorsal horn microglia displaying ranges of immunofluorescence (IF) intensity values of P2X₄R protein in individual microglia. Open bars, contralateral; filled bars, ipsilateral.

OX42 (ref. 14). We found that cells showing P2X₄R immunofluorescence were not double-labelled for MAP2 (Fig. 3a), NeuN (Fig. 3b) or GFAP (Fig. 3c). Instead, almost all P2X₄R-positive cells were double-labelled with OX42 (385 of 395 P2X₄R-positive cells examined; Fig. 3d), indicating that P2X₄Rs were expressed in microglia but not in neurons or astrocytes. OX42 recognizes complement receptor type 3 (CR3), the expression of which is greatly increased in hyperactive compared with resting microglia¹⁵. We found that OX42 labelling was greater in the dorsal horn ipsilateral to the nerve injury (Fig. 3e), whereas OX42 labelling in the dorsal horn was low bilaterally in sham-operated animals (not shown). OX42-positive cells were more numerous (Fig. 3e, f) and displayed hypertrophic morphology (Fig. 3g) in the dorsal horn on the side of the nerve injury in comparison with the contralateral side (Fig. 3j). These results indicate that nerve injury induced a switch from the resting to the hyperactive phenotype in the population of microglia in the dorsal horn. We found that cells expressing P2X₄R immunofluorescence in the ipsilateral dorsal horn had high levels of OX42 labelling (Fig. 3g–i) in comparison with the contralateral side (Fig. 3j–l). The mean intensity of P2X₄R immunofluorescence per OX42-positive cell was on average 5.4-fold higher in the ipsilateral dorsal horn than in the contralateral dorsal horn ($P < 0.001$, Fig. 3m) and the distribution of P2X₄R immunofluorescence intensities per OX42-positive cell was skewed to the right (Fig. 3n). We therefore concluded that in the dorsal horn after nerve injury hyperactive microglia are the cell type that expresses P2X₄R and that the P2X₄R expression is markedly increased in individual hyperactive microglia. Moreover, we observed that in primary culture, microglia that show the hyperactive phenotype express functional P2X₄Rs that are activated by 50 μ M ATP and inhibited by 10 μ M TNP-ATP but not by 10 μ M PPADS (see Supplementary Figure and Supplementary Information 1).

Because we find that the expression of P2X₄R is markedly

upregulated in individual microglia after nerve injury, the results above imply that tactile allodynia after nerve injury is critically dependent on functional P2X₄Rs in hyperactive microglia in the dorsal horn. We therefore predicted that suppressing the expression of P2X₄R should prevent tactile allodynia after nerve injury. We tested this by means of intrathecal treatment with an antisense oligodeoxynucleotide (ODN) targeting P2X₄R or with a mismatch ODN as a control. The animals were treated for 7 days, beginning on the day of the nerve lesion. We found that the nerve-injury-induced decrease in paw withdrawal threshold was significantly less in animals treated with P2X₄R antisense ODN (5 nmol, $n = 11$ animals) than in animals treated with mismatch ODN (5 nmol, $n = 10$ animals) ($P < 0.01$, Fig. 4a); the paw withdrawal threshold in animals treated with mismatch ODN was not different from that of untreated controls (Fig. 1a, $P > 0.48$). The paw withdrawal threshold in animals treated with P2X₄R antisense ODN recovered by $47.4 \pm 8.3\%$ towards the baseline before nerve injury (Fig. 4b). Furthermore, we found that the level of P2X₄R protein in homogenates from the spinal cord of antisense-ODN-treated rats ($n = 5$ animals) was $32.0 \pm 4.8\%$ less than that of mismatch-ODN-treated rats ($n = 4$ animals; $P < 0.01$, Fig. 4b). In addition, the immunofluorescence intensity of P2X₄R protein in individual microglia in the dorsal horn of antisense-ODN-treated rats ($n = 116$ microglial cells) was significantly lower than in those of mismatch-ODN-treated rats ($n = 115$ microglial cells; $37.1 \pm 0.2\%$ decrease, $P < 0.001$; Fig. 4b). However, we found that there was no difference in the number of microglia in the dorsal horn between rats treated with antisense ODN and those treated with mismatch ODN (Fig. 4c), nor was there a difference in OX42 immunofluorescence per individual microglial cell. In addition, microglia maintained a hypertrophic morphology in animals treated with P2X₄R antisense ODN (Fig. 4c). These results indicate that intrathecal treatment with P2X₄R antisense ODN suppressed both the tactile allodynia

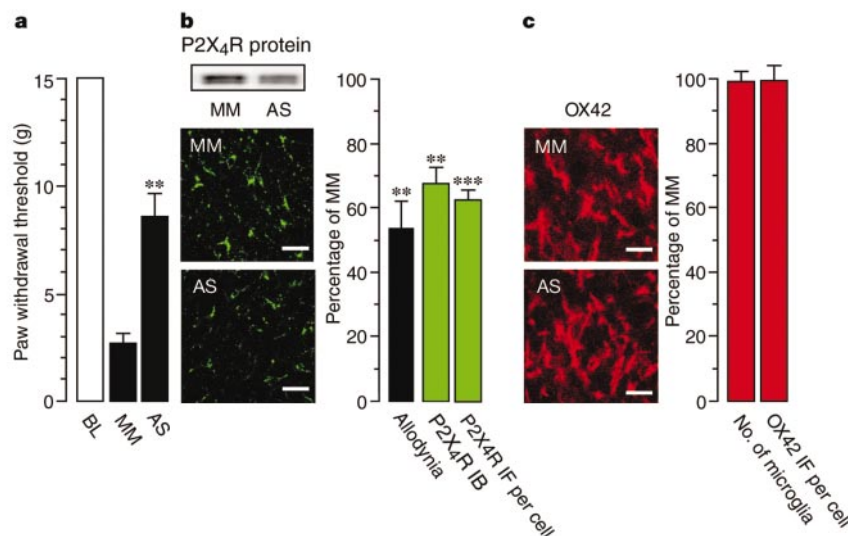


Figure 4 P2X₄R antisense ODN suppresses the development of tactile allodynia caused by injury to the L5 spinal nerve. Rats were injected intrathecally with antisense ODN (5 nmol) or mismatch ODN (5 nmol) once a day for 7 days. **a**, Paw withdrawal threshold (mean $\pm$ s.e.m.) of tactile stimulation to the hindpaw ipsilateral to the nerve injury. BL, baseline before nerve injury; MM, animals ($n = 10$) treated with mismatch ODN; AS, animals ($n = 11$) treated with antisense ODN (two asterisks, $P < 0.01$). **b**, Western blot (upper) and immunocytochemical (lower) analyses of P2X₄R protein in the spinal cord and in individual microglia in the dorsal horn, respectively, of rats treated with antisense ODN (AS) and mismatch ODN (MM). Scale bars, 20 μ m. Results from animals treated with antisense ODN were normalized to the values of the group treated with mismatch ODN

(mean $\pm$ s.e.m.). The behavioural effect of antisense ODN was converted into the percentage of tactile allodynia, which was calculated by the formula: tactile allodynia (%) = $100 \times (15.1 \text{ g} - \text{value of antisense ODN}) / (15.1 \text{ g} - \text{value of mismatch ODN})$ (two asterisks, $P < 0.01$). IB, immunoblot in homogenates from the spinal cord (two asterisks, $P < 0.01$); IF, immunofluorescence intensity in individual microglia (three asterisks, $P < 0.001$). **c**, Numbers of microglia and intensity of OX42 immunofluorescence in individual microglia in the dorsal spinal cord of rats treated with antisense ODN (AS) and mismatch ODN (MM). Scale bars, 10 μ m. Results were normalized to the values of the group treated with mismatch ODN (mean $\pm$ s.e.m.).

and the increase in P2X₄R expression after nerve injury, but treatment with P2X₄R antisense ODN did not suppress OX42 labelling nor prevent the switch to the hyperactive phenotype of the microglia in the dorsal horn.

We next investigated whether activation of P2X₄R in hyperactive microglia is sufficient to produce tactile allodynia. We tested the paw withdrawal threshold in normal rats after intrathecal injection of microglia that had been grown in primary culture (Fig. 5a) and therefore expressed P2X₄R and high levels of OX42. The microglia were preincubated for 1 h with PBS without or with ATP (50 μ M) to activate P2X₄R preferentially but not P2X₇R (see Supplementary Figure and Supplementary Information 1). We found that the paw withdrawal threshold decreased markedly after the intrathecal administration of ATP-stimulated microglia ($n = 7$ animals; Fig. 5b). In contrast, the paw withdrawal threshold was unaffected by intrathecal administration of PBS-treated microglia ($n = 5$ animals), PBS alone ($n = 4$ animals) or ATP (50 μ M) alone ($n = 4$ animals; Fig. 5b). The decrease in paw withdrawal threshold was prevented by including TNP-ATP (10 μ M) together with ATP in the preincubation ($n = 5$ animals; Fig. 5b). We therefore concluded

that P2X₄R-stimulated microglia were sufficient to produce tactile allodynia in otherwise naive rats.

The decrease in paw withdrawal threshold developed progressively over a 5-h period after administering the ATP-stimulated microglia (Fig. 5b); we therefore wondered whether stimulation of P2X₄R was required at the time when the allodynia was observed. We investigated this by intrathecal injection of TNP-ATP (30 nmol) 5 h after the ATP-stimulated microglia had been administered ($n = 5$ animals; Fig. 5c). We found that the paw withdrawal threshold increased gradually after the injection of TNP-ATP, peaking 45–75 min after the injection (45 and 60 min, $P < 0.05$; 75 min, $P < 0.01$ compared with before TNP-ATP). The paw withdrawal threshold at the peak of the effect of TNP-ATP was not different from the baseline before intrathecal injection of the ATP-stimulated microglia, indicating that TNP-ATP reversed the allodynia caused by the ATP-stimulated microglia. By 90 min after injection of TNP-ATP the paw withdrawal threshold had decreased significantly, implying a reversal of the effect of TNP-ATP and recovery of the tactile allodynia induced by administering the ATP-stimulated microglia. Taking these results together we concluded that stimulation of P2X₄R is required when tactile allodynia caused by ATP-stimulated microglia is observed, and this tactile allodynia is therefore like that caused by peripheral nerve injury. Furthermore, these results indicate that stimulation of microglia by P2X₄R is not only sufficient to induce tactile allodynia but also is necessary for maintaining the allodynia.

Here we show that the activation of P2X₄R induced in spinal cord microglia is essential for tactile allodynia after peripheral nerve injury. Tactile allodynia was reversed rapidly by pharmacological blockade of these receptors, implying that nerve-injury-induced hypersensitivity to pain depends on ongoing signalling by means of P2X₄R, probably activated by ATP, which might be released from primary sensory terminals^{16–18}, dorsal horn neurons^{16,19,20}, or from dorsal horn astrocytes²¹ (see also Supplementary Information 2). As a consequence of peripheral nerve injury, microglia in the spinal dorsal horn are converted to the hyperactive phenotype and have a markedly increased expression of P2X₄R. As we show that in hyperactive microglia activating P2X₄R provokes the entry of Ca²⁺ and that administering P2X₄R-stimulated hyperactive microglia produces tactile allodynia in normal rats, pain hypersensitivity might result from a release from these microglia of bioactive factors, such as cytokines^{22–24}, that enhance synaptic transmission in spinal pain pathways. Preventing the upregulation of P2X₄R expression in spinal microglia and inhibiting these receptors pharmacologically therefore represent novel therapeutic approaches for treating pain hypersensitivity caused by nerve damage, for which there is currently no effective therapy. Because a pharmacological blockade of these receptors did not affect pain responses in naive animals, a predicted therapeutic benefit of interfering with P2X₄R is that normal pain sensitivity would be unaffected. In this study we have discovered a crucial role for hyperactive microglia in a model of neuropathic pain, a common pathological condition of the central nervous system. Hyperactive microglia are considered crucial for the pathogenesis of various other pathological conditions of the central nervous system, such as neurodegenerative disorders and stroke^{25–27}. If P2X₄R are increased in microglia in these conditions then they could represent a novel therapeutic target in disorders of the central nervous system in addition to hypersensitivity to nerve-injury-induced pain. □

Methods

Behavioural studies

We used the spinal nerve injury model⁹ with some modifications: a unilateral L5 spinal nerve of male Wistar rats was tightly ligated and cut just distal to the ligature. The mechanical allodynia was assessed in naive and nerve-injured rats by using calibrated von Frey filaments (0.4–15.1 g). Rats were injected intrathecally with the P2XR antagonists TNP-ATP and PPADS (Sigma-RBI), and antisense and mismatch ODN. For the

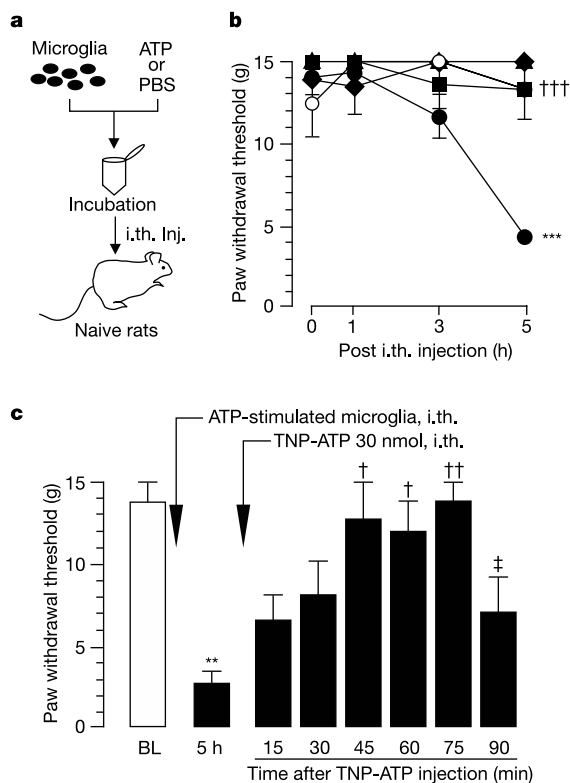


Figure 5 Spinal administration of ATP-stimulated microglia in normal rats produces tactile allodynia that depends on P2X₄R. **a**, Experimental protocol. i.th. inj., intrathecal injection. **b**, Paw withdrawal threshold (mean $\pm$ s.e.m.) of tactile stimulation after intrathecal administration of PBS (open circles), 50 μ M ATP (triangles) and microglia preincubated for 1 h with PBS (diamonds) or 50 μ M ATP (filled circles; three asterisks, $P < 0.001$ compared with PBS-treated microglia group). TNP-ATP (10 μ M; squares) was pretreated with microglia 10 min before the application of ATP (three daggers, $P < 0.001$ compared with ATP-stimulated microglia-treated group). **c**, Reversal, by intrathecal administration of TNP-ATP (30 nmol) 5 h after the injection of microglia, of tactile allodynia caused by the ATP-stimulated microglia (two asterisks, $P < 0.01$ compared with the preinjection baseline; dagger, $P < 0.05$; two daggers, $P < 0.01$ compared with the value at 5 h after the microglia injection; double dagger, $P < 0.05$ compared with the value at 75 min after the injection of TNP-ATP). Results are means $\pm$ s.e.m. of the paw withdrawal threshold (in grams).

experiments in Fig. 5, the cultured microglia (see Supplementary Figure) that had been preincubated with or without ATP (50 μ M) were injected intrathecally in normal rats (see Supplementary Methods for full details).

Immunohistochemistry

Transverse L5 spinal cord sections (30 μ m) were cut and processed for immunohistochemistry with anti-P2X₄R antibody (Alomone). Identification of the type of P2X₄R-positive cells was performed with the following markers: for microglia, OX42 (Chemicon) and iba1 (a gift from S. Kohsaka); for astrocytes, GFAP (Boehringer Mannheim); for spinal cord neurons, NeuN (Chemicon) and MAP2 (Chemicon). To assess immunofluorescence staining of cells quantitatively, we measured the immunofluorescence intensity of the P2X₄R or OX42 as the average pixel intensity within each cell (see also Supplementary Methods).

Western blotting

Western blot analysis of P2X₄R expression in the membrane fraction from L4–L6 spinal cord was performed with anti-P2X₄R polyclonal antibody (Oncogene) as described in detail in the Supplementary Methods.

Microglial culture

Rat primary cultured microglia were prepared in accordance with the method described previously²⁸. In brief, mixed glial culture was prepared from neonatal Wistar rats and maintained for 10–16 days in DMEM medium with 10% fetal bovine serum. Immediately before experiments, microglia were collected by a gentle shake as the floating cells over the mixed glial culture. The microglia were transferred to coverslips or to Eppendorf tubes for subsequent intrathecal administration.

Statistics

Statistical analyses of the results were made with Student's *t*-test, Student's paired *t*-test or the Mann–Whitney *U*-test.

Received 17 March; accepted 8 May 2003; doi:10.1038/nature01786.

1. Woolf, C. J. & Mannion, R. J. Neuropathic pain: Aetiology, symptoms, mechanisms, and management. *Lancet* **353**, 1959–1964 (1999).
2. Woolf, C. J. & Salter, M. W. Neuronal plasticity: Increasing the gain in pain. *Science* **288**, 1765–1769 (2000).
3. Bo, X., Zhang, Y., Nassar, M., Burnstock, G. & Schoepfer, R. A P2X purinoceptor cDNA conferring a novel pharmacological profile. *FEBS Lett.* **375**, 129–133 (1995).
4. Buell, G., Lewis, C., Collo, G., North, R. A. & Surprenant, A. An antagonist-insensitive P2X receptor expressed in epithelia and brain. *EMBO J.* **15**, 55–62 (1996).
5. Seguela, P., Haghighi, A., Soghomonian, J. J. & Cooper, E. A novel neuronal P2X ATP receptor ion channel with widespread distribution in the brain. *J. Neurosci.* **16**, 448–455 (1996).
6. Soto, F. *et al.* P2X₄: an ATP-activated ionotropic receptor cloned from rat brain. *Proc. Natl Acad. Sci. USA* **93**, 3684–3688 (1996).
7. Wang, C. Z., Namba, N., Gono, T., Inagaki, N. & Seino, S. Cloning and pharmacological characterization of a fourth P2X receptor subtype widely expressed in brain and peripheral tissues including various endocrine tissues. *Biochem. Biophys. Res. Commun.* **220**, 196–202 (1996).
8. Khakh, B. S. *et al.* International union of pharmacology. XXIV. Current status of the nomenclature and properties of P2X receptors and their subunits. *Pharmacol. Rev.* **53**, 107–118 (2001).
9. Kim, S. H. & Chung, J. M. An experimental model for peripheral neuropathy produced by segmental spinal nerve ligation in the rat. *Pain* **50**, 355–363 (1992).
10. Virginio, C., Robertson, G., Surprenant, A. & North, R. A. Trinitrophenyl-substituted nucleotides are potent antagonists selective for P2X₁, P2X₃, and heteromeric P2X_{2/3} receptors. *Mol. Pharmacol.* **53**, 969–973 (1998).
11. Tsuda, M., Ueno, S. & Inoue, K. Evidence for the involvement of spinal endogenous ATP and P2X receptors in nociceptive responses caused by formalin and capsaicin in mice. *Br. J. Pharmacol.* **128**, 1497–1504 (1999).
12. Zheng, J. H. & Chen, J. Modulatory roles of the adenosine triphosphate P2X-purinoceptor in generation of the persistent nociception induced by subcutaneous bee venom injection in the conscious rat. *Neurosci. Lett.* **278**, 41–44 (2000).
13. Tsuda, M., Ueno, S. & Inoue, K. *In vivo* pathway of thermal hyperalgesia by intrathecal administration of α,β -methylene ATP in mouse spinal cord: Involvement of the glutamate-NMDA receptor system. *Br. J. Pharmacol.* **127**, 449–456 (1999).
14. Honore, P. *et al.* Murine models of inflammatory, neuropathic and cancer pain each generates a unique set of neurochemical changes in the spinal cord and sensory neurons. *Neuroscience* **98**, 585–598 (2000).
15. Aldskogius, H. & Kozlova, E. N. Central neuron–glial and glial–glial interactions following axon injury. *Prog. Neurobiol.* **55**, 1–26 (1998).
16. Sawynok, J., Downie, J. W., Reid, A. R., Cahill, C. M. & White, T. D. ATP release from dorsal spinal cord synaptosomes: Characterization and neuronal origin. *Brain Res.* **610**, 32–38 (1993).
17. Li, P., Calejesan, A. A. & Zhuo, M. ATP P2X receptors and sensory synaptic transmission between primary afferent fibers and spinal dorsal horn neurons in rats. *J. Neurophysiol.* **80**, 3356–3360 (1998).
18. Nakatsuka, T. & Gu, J. G. ATP P2X receptor-mediated enhancement of glutamate release and evoked EPSCs in dorsal horn neurons of the rat spinal cord. *J. Neurosci.* **21**, 6522–6531 (2001).
19. Bardoni, R., Goldstein, P. A., Lee, C. J., Gu, J. G. & MacDermott, A. B. ATP P2X receptors mediate fast synaptic transmission in the dorsal horn of the rat spinal cord. *J. Neurosci.* **17**, 5297–5304 (1997).
20. Jo, Y. H. & Schlichter, R. Synaptic corelease of ATP and GABA in cultured spinal neurons. *Nature Neurosci.* **2**, 241–245 (1999).
21. Fam, S. R., Gallagher, C. J. & Salter, M. W. P2Y₁ purinoceptor-mediated Ca²⁺ signaling and Ca²⁺ wave propagation in dorsal spinal cord astrocytes. *J. Neurosci.* **20**, 2800–2808 (2000).
22. Inoue, K. Microglial activation by purines and pyrimidines. *Glia* **40**, 156–163 (2002).
23. Hanisch, U. K. Microglia as a source and target of cytokines. *Glia* **40**, 140–155 (2002).

24. Vitkovic, L., Bockaert, J. & Jacque, C. 'Inflammatory' cytokines: Neuromodulators in normal brain? *J. Neurochem.* **74**, 457–471 (2000).
25. Nakajima, K. & Kohsaka, S. Functional roles of microglia in the brain. *Neurosci. Res.* **17**, 187–203 (1993).
26. Carson, M. J. Microglia as liaisons between the immune and central nervous systems: Functional implications for multiple sclerosis. *Glia* **40**, 218–231 (2002).
27. Eikelenboom, P. *et al.* Neuroinflammation in Alzheimer's disease and prion disease. *Glia* **40**, 232–239 (2002).
28. Nakajima, K. *et al.* Identification of elastase as a secretory protease from cultured rat microglia. *J. Neurochem.* **58**, 1401–1408 (1992).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank J. Hicks for corrections to the manuscript. This work was supported by a Domestic Research Fellowship from the Japan Science and Technology Corporation, by a grant from the Uehara Memorial Foundation, partly by a grant from the Organization for Pharmaceutical Safety and Research, by a grant-in-aid for the scientific research from the Ministry of Education, Science, Sports, and Culture of Japan, and by a grant from the Japan Health Sciences Foundation. M.T. is supported by the Research Training Centre of the Hospital for Sick Children Research Institute. M.W.S. is an Investigator of the Canadian Institutes of Health Research.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to K.I. (inoue@nihs.go.jp).

A camelid antibody fragment inhibits the formation of amyloid fibrils by human lysozyme

Mireille Dumoulin¹, Alexander M. Last², Aline Desmyter^{3*}, Klaas Decanniere³, Denis Canet^{2*}, Göran Larsson¹, Andrew Spencer⁴, David B. Archer⁵, Jürgen Sasse⁶, Serge Muyldermans³, Lode Wyns³, Christina Redfield², André Matagne⁷, Carol V. Robinson¹ & Christopher M. Dobson¹

¹Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK

²Oxford Centre for Molecular Sciences, University of Oxford, South Parks Road, Oxford OX1 3QH, UK

³Department Ultrastructure, Vrije Universiteit Brussel, Pleinlaan 2, B-1050 Brussels, Belgium

⁴Institute of Food Research, Norwich Research Park, Colney, Norwich NR4 7UA, UK

⁵School of Life and Environmental Sciences, University of Nottingham, University Park, Nottingham NG7 2RD, UK

⁶Central Veterinary Research Laboratory, P.O. Box 597, Dubai, United Arab Emirates

⁷Laboratoire d'Enzymologie, Centre d'Ingénierie des Protéines, Institut de Chimie B6, Université de Liège, B-4000 Liège (Sart Tilman), Belgium

* Present addresses: Centre National de la Recherche Scientifique, 31 Chemin Joseph Aiguier, 13402 Marseille, France (A.D.); Gene Prot Inc., 2 Pré-de-la-Fontaine, 1217 Meyrin/GE, Switzerland (D.C.)

Amyloid diseases are characterized by an aberrant assembly of a specific protein or protein fragment into fibrils and plaques that are deposited in various organs and tissues^{1–3}, often with serious pathological consequences. Non-neuropathic systemic amyloidosis^{4–6} is associated with single point mutations in the gene coding for human lysozyme. Here we report that a single-domain fragment of a camelid antibody^{7–9} raised against wild-type human lysozyme inhibits the *in vitro* aggregation of its amyloidogenic variant, D67H. Our structural studies reveal that the epitope includes neither the site of mutation nor most residues in the region of the protein structure that is destabilized by the mutation. Instead, the binding of the antibody fragment achieves its effect by restoring the structural cooperativity characteristic

of the wild-type protein. This appears to occur at least in part through the transmission of long-range conformational effects to the interface between the two structural domains of the protein. Thus, reducing the ability of an amyloidogenic protein to form partly unfolded species can be an effective method of preventing its aggregation, suggesting approaches to the rational design of therapeutic agents directed against protein deposition diseases.

Four mutations in the lysozyme gene (I56T, F57I, W74R, and D67H) have so far been identified in connection with amyloid disease^{4–6}. The ability of the lysozyme variants studied in detail (I56T and D67H) to form amyloid deposits has been attributed primarily to their reduced stability relative to the wild-type protein, allowing transient unfolding of a local region of structure under physiologically relevant conditions^{10,11}. The formation of partly structured species through this locally cooperative unfolding process appears to be the critical event that triggers the aggregation process and ultimately leads to the formation of amyloid fibrils and the onset of disease¹¹.

The use of specific antibodies is a promising strategy for inhibiting and even reversing *in vitro* and *in vivo* fibril formation by amyloidogenic peptides or proteins^{12–14}. We focus here on the interaction of cAb-HuL6 (ref. 7), a fragment from a ‘heavy-chain’ camel antibody with high specificity for native human lysozyme and its amyloidogenic variants. Because antibodies derived from camels (camels, dromedaries and llamas) are devoid of the light chains of conventional antibodies, their antigen binding sites are limited to single domains, referred to as V_HHs⁸. Such domains have very favourable properties^{7–9} for biophysical studies, including small size and high solubility and stability.

To investigate the effects of the antibody fragment on fibril formation, conditions were selected (48 °C, pH 5.5, 3 M urea) under which the D67H variant readily aggregates *in vitro* while the antibody fragment remains stable and is still able to bind tightly to native lysozyme. In the absence of the antibody fragment, the data show that the D67H variant readily aggregates (Fig. 1a). The kinetics follow a sigmoidal curve consistent with the nucleation-dependent growth model that has been suggested as a common mechanism of fibril formation¹⁵. After incubation for 6 h, the protein solution was passed through a 0.22 µm filter and examined by SDS–polyacrylamide gel electrophoresis (SDS–PAGE). No protein could be detected on the gel (Fig. 1b, lane 4), indicating that virtually all the D67H protein had aggregated to form species too large to pass through the filter; the slight increase in light scattering observed over longer incubation times (6–15 h, Fig. 1a) is likely to result from the organization of the aggregates into higher-order structures. In contrast, when wild-type lysozyme was incubated under the same conditions, it remained almost completely soluble even after 20 h (Fig. 1a, b).

The aggregated species formed by the D67H variant under the conditions described above bind Congo red, producing the characteristic red shift in the absorption spectrum of the dye typically observed during amyloid fibril formation¹⁶. The presence of large quantities of fibrillar material was unambiguously confirmed by negative-stain transmission electron microscopy (TEM) (Fig. 1c). The structures observed include some well-resolved isolated fibrils (approximately 6 nm in diameter) although most fibrils appear to be incorporated into large bundles (100–400 nm in diameter). Such structures have previously been observed with a variety of other amyloidogenic proteins¹⁷, including the I56T variant of human lysozyme after incubation at pH 2 and 37 °C (ref. 18). Moreover, the X-ray fibre diffraction pattern of the fibrils (Fig. 1d) shows a dominant meridional reflection at 4.7 Å and an equatorial reflection at 10.4 Å, features characteristic of amyloid fibrils as a consequence of their cross-β structure¹⁹.

When the D67H protein was incubated in the presence of an equimolar amount of the antibody fragment, no significant changes in light scattering were detectable after 6 h of incubation (Fig. 1a),

and after passage of the solution through a 0.22 µm filter, SDS–PAGE confirmed that virtually all of the D67H protein incubated with the antibody fragment remained in solution (Fig. 1b, lane 6). Finally, TEM analysis showed no evidence for the presence of amyloid fibrils or other aggregates. Thus, in the presence of an equimolar quantity of antibody fragment, the D67H variant behaves very similarly to the wild-type protein in the absence of the antibody fragment. Moreover, the rate of aggregation of the D67H protein in the presence of the antibody fragment can be estimated to be reduced by a factor of at least 100 relative to that in its absence. In the presence of a sub-stoichiometric amount of the antibody fragment (a D67H:cAb-HuL6 ratio of 1:0.5), the initial rate of aggregation of the D67H variant was reduced by a factor of approximately 3 (Fig. 1a). Then, after about 50% of the protein had formed aggregates, the remainder aggregated even more slowly, presumably at a rate limited by the rate of dissociation of the D67H protein from the complex. These observations provide clear evidence that the binding of the antibody fragment to the variant protein is directly responsible for the dramatic decrease in its rate of aggregation.

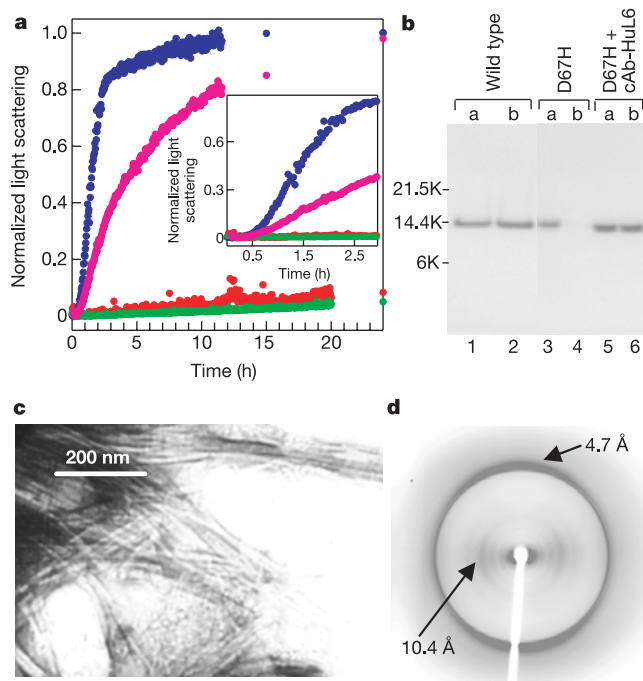


Figure 1 Effect of cAb-HuL6 on the aggregation of D67H lysozyme. **a**, Time course of the aggregation of D67H lysozyme in the absence (blue) and presence of the antibody fragment (D67H:cAb-HuL6 1:1 (green), D67H:cAb-HuL6 1:0.5 (pink)) monitored by light scattering. Data are also shown for wild-type lysozyme in the absence of cAb-HuL6 (red). Inset, expansion to demonstrate the lag phase at the onset of the aggregation process. **b**, SDS–PAGE of protein solutions filtered through a 0.22 µm filter before incubation at 48 °C (series a) and after 6 h of incubation at 48 °C (series b). Note that the D67H lysozyme and the cAb-HuL6 have similar relative molecular masses (14,715 and 14,220, respectively) and thus they migrate in closely similar position in lanes 5 and 6. The comparable intensities of the bands before and after heating suggest that both proteins remain soluble after treatment; this result was confirmed by measuring the protein concentration in the supernatant, using the BCA assay. **c**, Representative image of fibrils formed from the D67H variant produced by TEM. **d**, X-ray fibre diffraction pattern of the fibrils showing a prominent meridional reflection at 4.7 Å and an equatorial reflection at 10.4 Å, features typical of amyloid fibrils as a result of their cross-β structure. The additional equatorial reflections observed at 7.8 Å and 13.8 Å may arise from the superposition of interference functions as a consequence of the close packing of protofilaments¹⁹.

Thermal denaturation experiments followed by far-ultraviolet circular dichroism measurements under the buffer conditions in which the uncomplexed D67H variant forms fibrils indicate that the temperature of the mid-point of unfolding of the D67H variant ($T_m = 55 \pm 1^\circ\text{C}$) is reduced by about 10°C compared with that of the wild-type protein, ($T_m = 65 \pm 1^\circ\text{C}$). In the presence of an equimolar amount of the antibody fragment, however, the stability of the variant protein is raised by about 15°C ($T_m = 70 \pm 1^\circ\text{C}$). Within the complex, therefore, the thermal stability of the D67H lysozyme is even greater than that of the wild-type protein in the absence of the antibody fragment. Furthermore, thermally induced denaturation of the D67H variant was observed to be associated with a significant increase in the fluorescence intensity of 8-anilino-1-naphthalene-sulphonic acid (ANS) at 475 nm (data not shown), indicating that a partly unfolded intermediate species with one or more hydrophobic regions exposed to the solvent is significantly populated during unfolding. A similar enhancement of ANS fluor-

escence was not observed, however, in the presence of the antibody fragment. These data suggest that in the presence of the antibody fragment the D67H variant unfolds in a single cooperative two-state transition.

Further insight into the dynamics and structural cooperativity of the amyloidogenic D67H variant in the presence of the antibody fragment was gained from hydrogen/deuterium exchange pulse-labelling experiments analysed by mass spectrometry¹¹. The bimodal mass distribution observed with the D67H protein as exchange takes place (Fig. 2a) is indicative of the transient cooperative unfolding of a local region of the structure shown previously to consist of the β -domain and the adjacent C-helix¹¹. The time course of the intensity of the peak corresponding to the lower mass species in Fig. 2a provides a direct measure of the opening rate of the cooperative fluctuation giving rise to exchange (Fig. 2a inset); the time constant (τ) of 15.3 ± 0.6 s obtained from this analysis is consistent with earlier data obtained at higher pH values¹¹. Most significantly, however, the peak that is observed at lower mass in the absence of the antibody fragment (coloured yellow in Fig. 2a) is not observed in the spectrum when an equimolar amount of the antibody fragment is present (Fig. 2b). Instead, the D67H protein gives rise to a single peak whose mass progressively decreases with time, an observation indicative of the gradual exchange of the various labile deuterons in the protein. Similar behaviour has been observed with the wild-type protein in the absence of the antibody fragment and results from amide hydrogens undergoing exchange through highly localized independent fluctuations of the structure¹¹.

These observations demonstrate that, in the presence of an equimolar amount of the antibody fragment, virtually none of the molecules of the D67H variant undergoes even a single locally cooperative unfolding event on the timescale of the experiment (about 1 h). This result indicates that the frequency of such fluctuations in the D67H variant is reduced by a factor of at least 2,000 as a result of binding to the antibody fragment. The presence

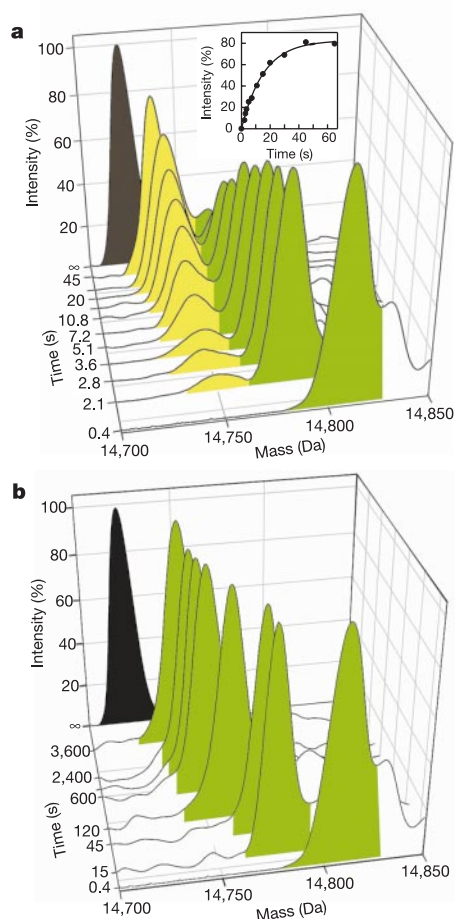


Figure 2 Electrospray mass spectra of D67H lysozyme in the absence (**a**) and presence (**b**) of an equimolar amount of cAb-HuL6. The peaks observed in spectra of control samples recorded after complete hydrogen–deuterium exchange are coloured black. The peaks coloured green arise from the gradual loss of deuterium during the course of the exchange reaction that occurs by an EX2 mechanism¹¹. The peaks coloured yellow are observed in the spectra of the D67H variant in the absence of the antibody fragment but not in its presence. They result from the locally cooperative unfolding event giving rise to exchange by an EX1 mechanism¹¹. Note the difference in the timescale between **a** and **b**. Inset in **a**, time course of the relative intensity of the lower mass species (peaks coloured yellow) observed for the unbound D67H variant (Fig. 2a). Fitting these data to an exponential function indicates that the time constant of the unfolding process is 15.3 ± 0.6 s.

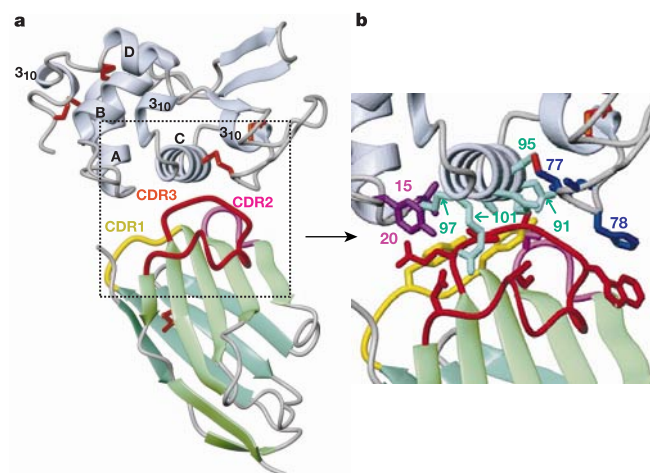


Figure 3 X-ray structure of wild-type human lysozyme complexed with cAb-HuL7. **a**, Ribbon representation of the complex. The β -strands in cAb-HuL6 are coloured green. The lysozyme molecule is shown in grey and the helices are labelled. The disulphide bridges are coloured orange. **b**, Enlarged view of the binding interface (that is, the area in the box in **a**). The side chains of residues constituting the epitope are shown in violet, light blue and dark blue for the α -domain, C-helix and β -domain of lysozyme, respectively. Those residues constituting the paratope are shown in yellow, pink and red, and are from the CDR1, CDR2 and CDR3 loops of the V_HH structure, respectively. The structure was produced by using MOLMOL (ref. 30).

of a stoichiometric amount of a V_HH that does not interact with lysozyme, cAb-R27, was found to have no effect on the dynamics of the D67H variant confirming that it is the specific binding of the cAb-HuL6 that is responsible for the inhibition of the partial unfolding event. The data from these various experiments, therefore, provide evidence that the formation of partly unstructured species, resulting from the locally cooperative unfolding of the β -domain and the C-helix is the critical event that triggers the aggregation process in the absence of the antibody fragment. Moreover, the results show that binding to the antibody fragment not only increases the overall stability of the D67H variant to a value even greater than that of the wild-type protein, but also reverses the loss of global structural cooperativity that results from the mutation.

To investigate in greater detail the structural basis for the efficacy of the antibody fragment as an inhibitor of the formation of amyloid fibrils, the complex with wild-type lysozyme was crystallized and the structure solved at a resolution of 1.8 Å by using X-ray diffraction (Fig. 3). Comparison of this structure with those available for unbound lysozyme reveals that no significant conformational changes result from the complexation of the protein to the antibody fragment; the C α r.m.s. deviations of the lysozyme molecule in the complex from that of the uncomplexed species (pdb entries 1LZ1, 1REX and 1JWR) are all less than 0.4 Å. The structure of the complex reveals that the epitope consists of 14 residues of the lysozyme molecule and encompasses residues located in the loop between the A and B helices in the α -domain (L15, G16, Y20), in the

long loop within the β -domain (A76, C77, H78, L79) and in the C-helix (A90, D91, A94, C95, K97, R 98, R101) (Fig. 3); binding of the antibody fragment results in a surface coverage in these regions of about 100 Å², 165 Å² and 305 Å², respectively. Thus, the C-helix contributes about half of the total surface area of the enzyme that is bound to the antibody fragment.

We performed NMR experiments to characterize further the interaction in solution between the wild-type and variant lysozyme molecules and the antibody fragment. Comparison of the ¹⁵N-¹H heteronuclear single quantum coherence (HSQC) NMR spectra of the unbound proteins, uniformly labelled with ¹⁵N, with spectra of the proteins complexed to an unlabelled antibody fragment allowed the binding region to be probed in a site-specific manner by analysis of the chemical shift perturbations of the resonances of the amide groups of individual residues. This analysis reveals that the resonances of approximately the same number of residues in the spectra of the wild-type lysozyme and the D67H variant (31 and 29, respectively) exhibit a significant chemical shift perturbation upon binding to the antibody fragment in solution (Fig. 4). Resonances of all the lysozyme residues found to be in direct contact with the antibody fragment in the X-ray structure were observed to be significantly shifted relative to their positions in the uncomplexed protein. Most of the additional residues with perturbed chemical shifts are located in close proximity to those that are in direct contact with the antibody fragment. Moreover, it is evident that the epitope is essentially identical in the D67H variant as in wild-type lysozyme. This last finding is consistent with the very

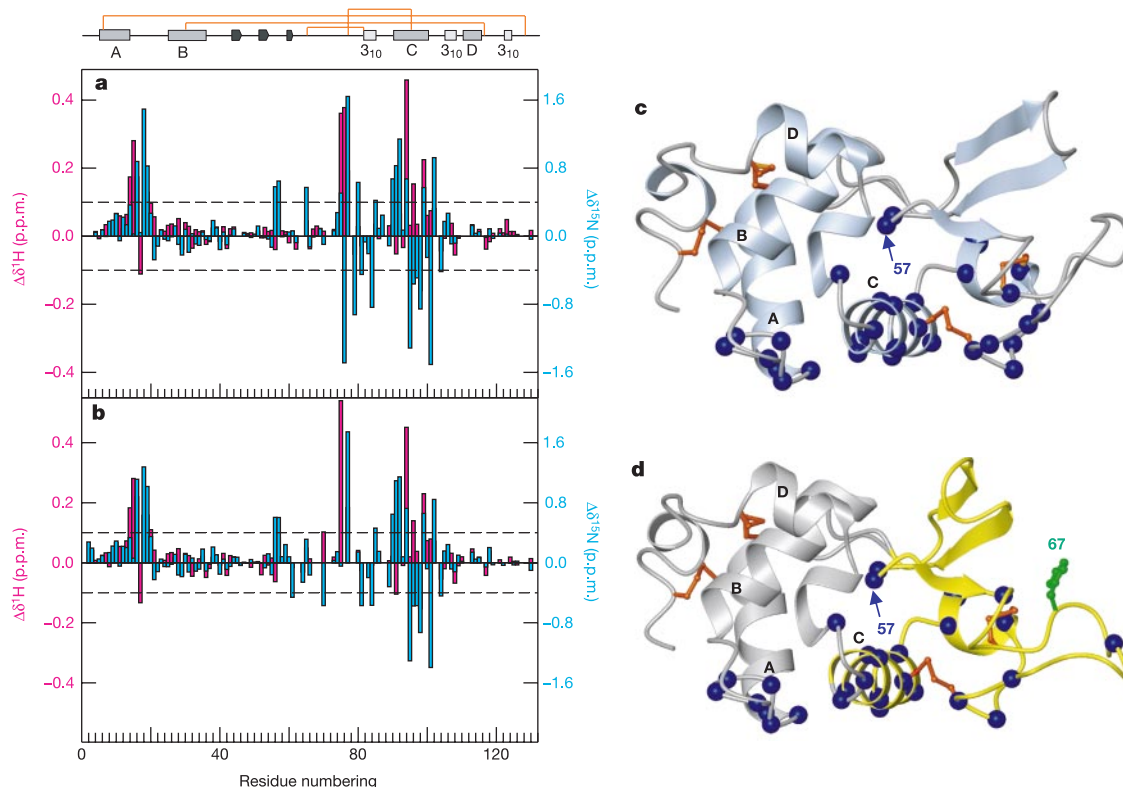


Figure 4 Chemical shift perturbations induced by the binding of the cAb-HuL6 to wild-type lysozyme (**a**) and the D67H variant (**b**). The pink and blue bars indicate the chemical shift perturbations of the ¹H and ¹⁵N resonance, respectively; residues experiencing a chemical shift change of not less than ± 0.4 p.p.m. for ¹⁵N or not less than ± 0.1 p.p.m. for ¹H resonances are considered to be significantly affected by the binding of the antibody fragment. The C α atoms of these residues are shown in a blue space-filling representation on the ribbon diagram (**c**, wild-type lysozyme; **d**, D67H variant). The peaks corresponding to residues 76, 78 and 79 in the D67H-cAbHuL6 complex and to residue

78 in the wild-type protein complex have not been assigned; therefore shift differences for these peaks are not shown in the histograms. No unassigned peaks are visible in the spectra of either complex within 0.1 p.p.m. (¹H) and 0.4 p.p.m. (¹⁵N) of the peak positions of residue 78 in the spectra of the unbound proteins. This finding strongly suggests that large shift perturbations are observed for this residue in both complexes. In **d**, the region of the D67H molecule that has been found to unfold transiently in a locally cooperative manner¹¹ is coloured yellow, and the amyloidogenic mutation D67H is indicated in green.

similar values of the dissociation constants for binding of the antibody fragment to wild-type lysozyme ($K_D = 0.7$ nM (ref. 7)) and the D67H variant ($K_D = 1.2$ nM, this work).

The crystal structure of the complex indicates that the antibody fragment binds to the C-terminal region of the long loop of the β -domain and to the C-helix of the α -domain of lysozyme. Remarkably, these are two of the elements of secondary structure that are destabilized by the D67H amyloidogenic mutation as a result of the disruption of the interface between the α - and β -domains^{10,11} (Fig. 4d). This result therefore provides a ready explanation for the ability of the antibody fragment to prevent the partial unfolding and subsequent aggregation of the D67H variant. More detailed analysis, however, reveals that only 11 of the nearly 60 residues involved in the transient unfolding of the D67H variant are in direct contact with the antibody fragment. Thus, the effects of binding are not simply to mask the entire region of the protein destabilized by the mutation and hence to prevent its unfolding from the remainder of the structure. Rather, it appears that the binding of the antibody fragment restores the global cooperativity of the lysozyme structure that is disrupted by the D67H mutation by a more subtle mechanism. It is therefore of particular interest that perturbations to the NMR chemical shifts, reflecting small changes in the protein structure, include the amide resonances of residues 56 and 57, both of which are far from the site of mutation and from the binding interface with the antibody. In the crystal structure of the D67H variant it was found that the mutation disrupts the hydrogen bond network present in the β -domain of the wild-type protein, and that the resulting small conformational changes are transmitted to the region of the structure where residue 56 is located¹⁰. Moreover, residues 56 and 57 are themselves the locations of two of the other well-defined pathogenic mutations of lysozyme that give rise to amyloid disease^{4,6}.

As the result of the increasingly serious impact of amyloid diseases such as Alzheimer's and Parkinson's on the ageing populations of the developed world, a wide range of strategies for the prevention or treatment of these diseases is being very actively explored^{20–24}. One of the most attractive therapeutic options is to prevent the accumulation of aggregation-prone species by inhibiting the processes that lead to their formation²³. Studies of transthyretin, for example, have shown that small molecules mimicking the binding of natural ligands are highly promising therapeutic agents because they stabilize the native state of the protein, in this case a tetramer, by binding at the interface between the subunits and preventing their dissociation under physiological conditions^{25,26}. We have shown here that a specific antibody fragment can achieve a similar effect in reducing the concentration of a partly unfolded aggregation-prone species, but in this case by stabilizing the interactions between distinct sub-structures of a monomeric protein. Unlike the situation with transthyretin, stabilization and inhibition of lysozyme aggregation does not occur as a result of binding at a site that is normally occupied by a natural ligand. Rather, it occurs through the restoration of the global cooperativity of the native structure, disrupted by the amyloidogenic mutation that is a consequence of the binding interaction. The region of the protein surface that forms the epitope of the antibody fragment does not include the site of the disruptive mutation; the antibody fragment thus acts through transmitted conformational changes in a manner that is in some respect analogous to the way in which binding of an effector molecule can regulate an allosteric protein²⁷. This suggests that exploration of a series of antibodies raised against a given protein antigen could well allow the discovery of species able to overcome the effects of a wide range of pathogenic mutations, or indeed to stabilize a wild-type protein without disruption of the binding site or sites that are essential to its function. The properties of single domain fragments of camelid antibodies may make humanized versions of these molecules particularly favourable for such a purpose⁹. □

Methods

Proteins

Wild-type human lysozyme and its D67H variant, including uniformly [¹⁵N]proteins, cAb-R2 and the cAb-HuL6 were expressed and purified as described previously^{7,28}. The affinity of cAb-HuL6 for the two forms of lysozyme was determined by using surface plasmon resonance⁷.

Thermal unfolding monitored by circular dichroism and fluorescence measurements

Thermal unfolding experiments were done in 0.1 M citrate buffer pH 5.5 containing 3 M urea, by raising the temperature from 5 to 75–95 °C at a rate of 0.5 °C min^{−1}. The far ultraviolet circular dichroism signal was monitored at 228 nm by using a Jasco J-810 spectropolarimeter with a 0.1-cm path-length cell using lysozyme concentrations of 13.5 μ M. At this wavelength the signal arising from the antibody fragment is very low compared with that of lysozyme⁷. The changes in signal amplitude as a result of unfolding of the antibody fragment are therefore negligible, making it possible to monitor specifically the unfolding of lysozyme in the presence of cAb-HuL6. ANS-binding fluorescence emission data were recorded on a Cary Eclipse spectrofluorimeter. The excitation and emission wavelengths were 350 nm and 475 nm, respectively, and the slit widths were 5 nm. The lysozyme concentration was 2.5 μ M in a 1-cm path-length cuvette and the final ANS concentration was 160 μ M. For the experiments in the presence of the antibody fragment, equimolar amounts of cAb-HuL6 and D67H lysozyme were mixed immediately before the experiment.

Aggregation monitored by right-angle light scattering

Stock solutions of the proteins were prepared from lyophilized material, passed through a 0.22 μ m filter, diluted to the desired final concentration into 0.1 M citrate buffer pH 5.5 containing 3 M urea and incubated at 48 °C with stirring in a Cary Eclipse spectrofluorimeter. The excitation wavelength was 500 nm and changes in fluorescence emission were monitored at 500 nm with slit widths of 5 nm. Aliquots of the samples after 6 h of incubation were centrifuged for 15 min at 15,000 g. Pellets were resuspended in 50 μ l of water and used for X-ray diffraction measurements. The supernatant was passed through a 0.22 μ m filter and its protein concentration was analysed by SDS–PAGE (NuPAGE, 4–12% Bis-Tris, Invitrogen) and the bicinchoninic acid (BCA) assay (Pierce).

Congo red binding

Congo red binding assays were performed as described previously¹⁶ by using a Cary 4 UV–Vis spectrophotometer.

Electron microscopy

Samples were applied to Formvar-coated nickel grids, stained with 2% (w/v) uranyl acetate solution and viewed in a Phillips CEM100 transmission electron microscope operating at 80 kV.

X-ray fibre diffraction

Samples were prepared by air-drying fibril preparations between two waxed-filled capillary ends, aligned in the X-ray beam. Diffraction images were collected with a Rigaku Cu K α rotating-anode source (wavelength 1.54 Å) and an R-Axis IV image-plate X-ray detector.

Mass spectrometry

Pulse labelling experiments, which involved the replacement of deuterium atoms with hydrogen atoms from the solvent H₂O, were done with D67H lysozyme that had initially been deuterated at all the labile exchangeable sites and a Bio-Logic QFM5 mixer, as described previously¹¹. The samples that were allowed to exchange for more than 2 min were prepared by manual mixing. For the experiments in the presence of an equimolar amount of cAb-HuL6, deuterated D67H lysozyme and protonated cAb-HuL6 were mixed immediately before the experiment was performed. Exchange was done at pH 8.0 (100 mM ammonium/formic acid in H₂O) and 37 °C for various lengths of time and then quenched by decreasing the temperature and pH. The samples were electrosprayed at the base pressure of the Platform mass spectrometer (Micromass) and at a cone voltage of 150 V, where the complex between the antibody fragment and the lysozyme was found to dissociate. Mass spectra shown in Fig. 2 represent the convolution of the +8, +9, +10 charge states with minimal smoothing and converted to a mass scale.

Crystal structure determination

Crystals of cAb-HuL6 complexed with wild-type lysozyme were grown by the hanging drop vapour diffusion method at 20 °C. Drops initially contained equal volumes of protein (0.2 mM in buffer) and reservoir buffer (20% w/v PEG 4000 and 0.2 M imidazole malate, pH 6.0). Details of the data collection and processing will be described elsewhere (Decanniere *et al.*, manuscript in preparation). The atomic coordinates have been deposited in the Protein Data Bank (accession number 1OP9).

NMR spectroscopy

Two-dimensional gradient-enhanced ¹⁵N-¹H HSQC spectra of unbound wild-type and D67H lysozyme, and of the proteins complexed with cAb-HuL6, were collected at 750 MHz at 35 °C. The samples of the complexes were made up with ~0.6 and ~0.4 mM ¹⁵N-labelled wild-type and D67H lysozyme, respectively, and with a ~20% excess of unlabelled cAb-HuL6, to ensure all the lysozyme was bound in the complex, at pH 6.5 in 20 mM phosphate buffer in 95% H₂O/5% D₂O. The HSQC spectra were collected with 64 transients per t_1 increment, with 128 and 1024 complex points in t_1 and t_2 , and with sweep widths of 10,582 and 2,273 Hz in F_2 and F_1 . The ¹H and ¹⁵N resonances of wild-type

lysozyme in the complex were assigned by using a ^{15}N -edited three-dimensional nuclear Overhauser enhancement (NOE) spectroscopy—HSQC experiment. This was collected with eight transients per increment, with 128, 28 and 512 complex points in t_1 (^1H), t_2 (^{15}N) and t_3 (^1H), and with acquisition times of 14.2, 12.3 and 48.4 ms in t_1 , t_2 and t_3 . The observed NOE data for bound lysozyme were interpreted by using published assignments for the free protein²⁹. The resonances of D67H were assigned by comparison with the assigned spectrum of the unbound protein¹¹ and that of the wild-type complex.

Received 22 March; accepted 18 June 2003; doi:10.1038/nature01870.

1. Tan, S. Y. & Pepys, M. Amyloidosis. *Histopathology* **25**, 403–414 (1994).
2. Koo, E. H., Lansbury, P. T. Jr & Kelly, J. W. Amyloid diseases: Abnormal protein aggregation in neurodegeneration. *Proc. Natl Acad. Sci. USA* **96**, 9989–9990 (1999).
3. Dobson, C. M. The structural basis of protein folding and its links with human disease. *Phil. Trans. R. Soc. Lond. B* **356**, 133–145 (2001).
4. Pepys, M. B. *et al.* Human lysozyme gene mutations cause hereditary systemic amyloidosis. *Nature* **362**, 553–557 (1993).
5. Valleix, S. *et al.* Hereditary renal amyloidosis caused by a new variant lysozyme W64R in a French family. *Kidney Int.* **61**, 907–912 (2002).
6. Yazaki, M., Farrell, S. A. & Benson, M. D. A novel lysozyme mutation Phe57Ile associated with hereditary renal amyloidosis. *Kidney Int.* **63**, 1652–1657 (2003).
7. Dumoulin, M. *et al.* Single-domain antibody fragments with high conformational stability. *Protein Sci.* **11**, 500–515 (2002).
8. Hamers-Casterman, C. *et al.* Naturally occurring antibodies devoid of light chains. *Nature* **363**, 446–448 (1993).
9. Muyldermans, S. Single domain camel antibodies: current status. *J. Biotechnol.* **74**, 277–302 (2001).
10. Booth, D. R. *et al.* Instability, unfolding and aggregation of human lysozyme variants underlying amyloid fibrillogenesis. *Nature* **385**, 787–793 (1997).
11. Canet, D. *et al.* Local cooperativity in the unfolding of an amyloidogenic variant of human lysozyme. *Nature Struct. Biol.* **9**, 308–315 (2002).
12. Schenk, D. Amyloid-beta immunotherapy for Alzheimer's disease: the end of the beginning. *Nature Rev. Neurosci.* **3**, 824–828 (2002).
13. Peretz, D. *et al.* Antibodies inhibit prion propagation and clear cell cultures of prion infectivity. *Nature* **412**, 739–743 (2001).
14. White, A. R. *et al.* Monoclonal antibodies inhibit prion replication and delay the development of prion disease. *Nature* **422**, 80–83 (2003).
15. Harper, J. D. & Lansbury, P. T. Jr Models of amyloid seeding in Alzheimer's disease and scrapie: mechanistic truths and physiological consequences of the time-dependent solubility of amyloid proteins. *Annu. Rev. Biochem.* **66**, 385–407 (1997).
16. Klunk, W. E., Pettigrew, J. W. & Abraham, D. J. Two simple methods for quantifying low-affinity dye-substrate binding. *J. Histochem. Cytochem.* **37**, 1293–1297 (1989).
17. Colon, W. & Kelly, J. W. Partial denaturation of transthyretin is sufficient for amyloid fibril formation in vitro. *Biochemistry* **31**, 8654–8660 (1992).
18. Morozova-Roche, L. A. *et al.* Amyloid fibril formation and seeding by wild-type human lysozyme and its disease-related mutational variants. *J. Struct. Biol.* **130**, 339–351 (2000).
19. Sunde, M. *et al.* Common core structure of amyloid fibrils by synchrotron X-ray diffraction. *J. Mol. Biol.* **273**, 729–739 (1997).
20. Aguzzi, A., Glatzel, M., Montrasio, F., Prinz, M. & Heppner, F. L. Interventional strategies against prion diseases. *Nature Rev. Neurosci.* **2**, 745–749 (2001).
21. Wolfe, M. S. Therapeutic strategies for Alzheimer's disease. *Nature Rev. Drug Discov.* **1**, 859–866 (2002).
22. Pepys, M. B. *et al.* Targeted pharmacological depletion of serum amyloid P component for treatment of human amyloidosis. *Nature* **417**, 254–259 (2002).
23. Dobson, C. M. Protein folding and disease: a view from the first Horizon Symposium. *Nature Rev. Drug Discov.* **2**, 154–160 (2003).
24. May, B. C. *et al.* Potent inhibition of scrapie prion replication in cultured cells by bis-acridines. *Proc. Natl Acad. Sci. USA* **100**, 3416–3421 (2003).
25. McCammon, M. G. *et al.* Screening transthyretin amyloid fibril inhibitors. Characterization of novel multiprotein, multiligand complexes by mass spectrometry. *Structure (Cambridge)* **10**, 851–863 (2002).
26. Hammarstrom, P., Wiseman, R., Powers, E. T. & Kelly, J. W. Prevention of transthyretin amyloid disease by changing protein misfolding energetics. *Science* **299**, 713–716 (2003).
27. Branden, C. & Tooze, J. *Introduction to Protein Structure* (Garland, New York, 1999).
28. Spencer, A. *et al.* Expression, purification, and characterization of the recombinant calcium-binding equine lysozyme secreted by the filamentous fungus *Aspergillus niger*: Comparisons with the production of hen and human lysozymes. *Protein Expr. Purif.* **16**, 171–180 (1999).
29. Ohkubo, T., Taniyama, Y. & Kikuchi, M. ^1H and ^{15}N NMR study of human lysozyme. *J. Biochem. (Tokyo)* **110**, 1022–1029 (1991).
30. Koradi, R., Billeter, M. & Wuthrich, K. MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graph.* **14**, 51–55 (1996).

Acknowledgements The assistance of J. Zurdo, M. Krebs and B. Luisi for TEM and X-diffraction analysis of fibrils is gratefully acknowledged. We thank J.-M. Frère for many discussions. M.D. and D.C. were supported by a fellowship from the European Community. G.L. was supported by a fellowship from the Wenner-Gren Foundation. C.R. was supported by a BBSRC Advanced Research Fellowship. A.M. is a Research Associate of the FNRS, and was supported in part by a grant from the FRFC. C.V.R. is a Royal Society University Research Fellow. The research of C.M.D. is supported in part by a Programme Grant from the Wellcome Trust. This work was also supported by a BBSRC grant (to C.M.D., C.V.R. and D.B.A.) and by the Belgian Government through the PAL.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.M.D. (cmd44@cam.ac.uk).

Comparative analyses of multi-species sequences from targeted genomic regions

J. W. Thomas^{1*}, J. W. Touchman^{1,2*}, R. W. Blakesley^{1,2}, G. G. Bouffard^{1,2}, S. M. Beckstrom-Sternberg^{1,2}, E. H. Margulies¹, M. Blanchette³, A. C. Siepel³, P. J. Thomas², J. C. McDowell², B. Maskeri², N. F. Hansen², M. S. Schwartz³, R. J. Weber³, W. J. Kent³, D. Karolchik³, T. C. Bruen³, R. Bevan³, D. J. Cutler⁴, S. Schwartz⁵, L. Elnitski⁵, J. R. Idol¹, A. B. Prasad¹, S.-Q. Lee-Lin¹, V. V. B. Maduro¹, T. J. Summers¹, M. E. Portnoy¹, N. L. Dietrich², N. Akhter², K. Ayele², B. Benjamin², K. Cariaga², C. P. Brinkley², S. Y. Brooks², S. Granite², X. Guan², J. Gupta², P. Haghighi², S.-L. Ho², M. C. Huang², E. Karlins², P. L. Laric², R. Legaspi², M. J. Lim², Q. L. Maduro², C. A. Masiello², S. D. Mastrian², J. C. McCloskey², R. Pearson², S. Stantripop², E. E. Tiongsong², J. T. Tran², C. Tsurgeon², J. L. Vogt², M. A. Walker², K. D. Wetherby², L. S. Wiggins², A. C. Young², L.-H. Zhang², K. Osoegawa⁶, B. Zhu⁶, B. Zhao⁶, C. L. Shu⁶, P. J. De Jong⁶, C. E. Lawrence⁷, A. F. Smit⁸, A. Chakravarti⁴, D. Haussler^{3,9}, P. Green¹⁰, W. Miller⁵ & E. D. Green^{1,2}

¹Genome Technology Branch, National Human Genome Research Institute, and ²NIH Intramural Sequencing Center, National Institutes of Health, Bethesda, Maryland 20892, USA

³Center for Biomolecular Science and Engineering, University of California, Santa Cruz, California 95064, USA

⁴Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21287, USA

⁵Department of Computer Science and Engineering, The Pennsylvania State University, University Park, Pennsylvania 16802, USA

⁶Children's Hospital Oakland Research Institute, Oakland, California 94609, USA

⁷The Wadsworth Center for Laboratories and Research, New York State Department of Health, Albany, New York 12201, USA

⁸The Institute for Systems Biology, Seattle, Washington 98103, USA

⁹Howard Hughes Medical Institute, University of California, Santa Cruz, California 95064, USA

¹⁰Howard Hughes Medical Institute and Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA

* Present addresses: Department of Human Genetics, Emory University School of Medicine, Atlanta, Georgia 30322, USA (J.W.Th.); Translational Genomics Research Institute, Phoenix, Arizona 85004 and Department of Biology, Arizona State University, Tempe, Arizona 85287, USA (J.W.To.)

The systematic comparison of genomic sequences from different organisms represents a central focus of contemporary genome analysis. Comparative analyses of vertebrate sequences can identify coding^{1–6} and conserved non-coding^{4,6,7} regions, including regulatory elements^{8–10}, and provide insight into the forces that have rendered modern-day genomes⁶. As a complement to whole-genome sequencing efforts^{3,5,6}, we are sequencing and comparing targeted genomic regions in multiple, evolutionarily diverse vertebrates. Here we report the generation and analysis of over 12 megabases (Mb) of sequence from 12 species, all derived from the genomic region orthologous to a segment of about 1.8 Mb on human chromosome 7 containing ten genes, including the gene mutated in cystic fibrosis. These sequences show conservation reflecting both functional constraints and the neutral mutational events that shaped this genomic region. In particular, we identify substantial numbers of conserved non-coding segments beyond those previously identified experimentally, most of which are not detectable by pair-wise sequence comparisons alone. Analysis of transposable element insertions highlights the variation in genome dynamics among these species and confirms the placement of rodents as a sister group to the primates.

The NIH Intramural Sequencing Center (NISC) Comparative Sequencing Program aims to sequence and to analyse targeted genomic regions in multiple vertebrates. Our initial target is a genomic segment of about 1.8 Mb on human chromosome 7q31.3

containing the gene encoding the cystic fibrosis transmembrane conductance regulator¹¹ and nine other genes (referred to below as the 'greater *CFTR* region'). We sought to clone and to sequence the orthologous genomic segments in multiple other vertebrates (Table 1 and Methods). So far, our efforts have yielded more than 12 Mb of high-quality comparative sequence data (see Supplementary Information), over 95% of which has been finished to the standards established for human genome sequence¹². This represents the most diverse collection of large blocks of orthologous vertebrate sequence generated to date.

To identify regions of sequence conservation, we used blastz¹³ to construct pair-wise alignments of the sequences and MultiPip-Maker¹⁴ to show pair-wise percentage-identity plots of an annotated reference sequence against multiple query sequences (Fig. 1a and Supplementary Information). Alignments between the human sequence and the sequence of each of the other 12 species allowed the general patterns of conservation to be investigated. As expected, the fraction of sequence that can be aligned generally decreases with increasing evolutionary distance from humans (Fig. 1b). The only exceptions to this trend are mouse and rat, which, although considered to be closer to humans than to the other non-primate mammals included here¹⁵, have a lower fraction of sequence that can be aligned with the human sequence. This probably reflects a particularly high rate of sequence evolution, including large deletions, in the rodent lineage⁶.

For all species, reasonably consistent alignments are seen in coding exons (Fig. 1a, b). The human–fish alignments are largely limited to these, with only 14% of the aligned sequence outside coding exons; in fact, only two local human–fish alignments do not at least partially overlap a coding exon (5' to *CAV2* and within intron 4 of *CORTBP2*; see Supplementary Information). Almost a third (31.4%) of the human coding sequence does not align to fish sequence in the orthologous region. By contrast, the human–chicken alignments exclude less than 2% of the coding sequence, which is comparable to the alignments between human and the other mammals. Within the alignments, sequence divergence relative to human (measured as the percentage of single-nucleotide mismatches) varies from a low of 1.15% for chimpanzee to a high of 35.60% for zebrafish (see Supplementary Information).

Analyses of large-scale mutational events in this genomic region support mammalian phylogenies^{15,16} that place the rodents and primates as sister groups in one clade and the artiodactyls and carnivores as sister groups in another clade (Fig. 1c). Both of these groupings are confirmed by transposon insertions that are present in sister groups and absent from the other species. In particular, we found three clear examples of insertions that support the rodent–primate grouping (all *MLT1A0* elements) and three that support the artiodactyl–carnivore grouping (one *MLT1A0* and two *L1MA9*; see

Supplementary Information). In each case, insertions of the same transposon subtype, present in the same orientation and with the same target-site duplication, were identified at homologous positions in all members of sister groups.

Both *MLT1A0* and *L1MA9* elements are thought to have been

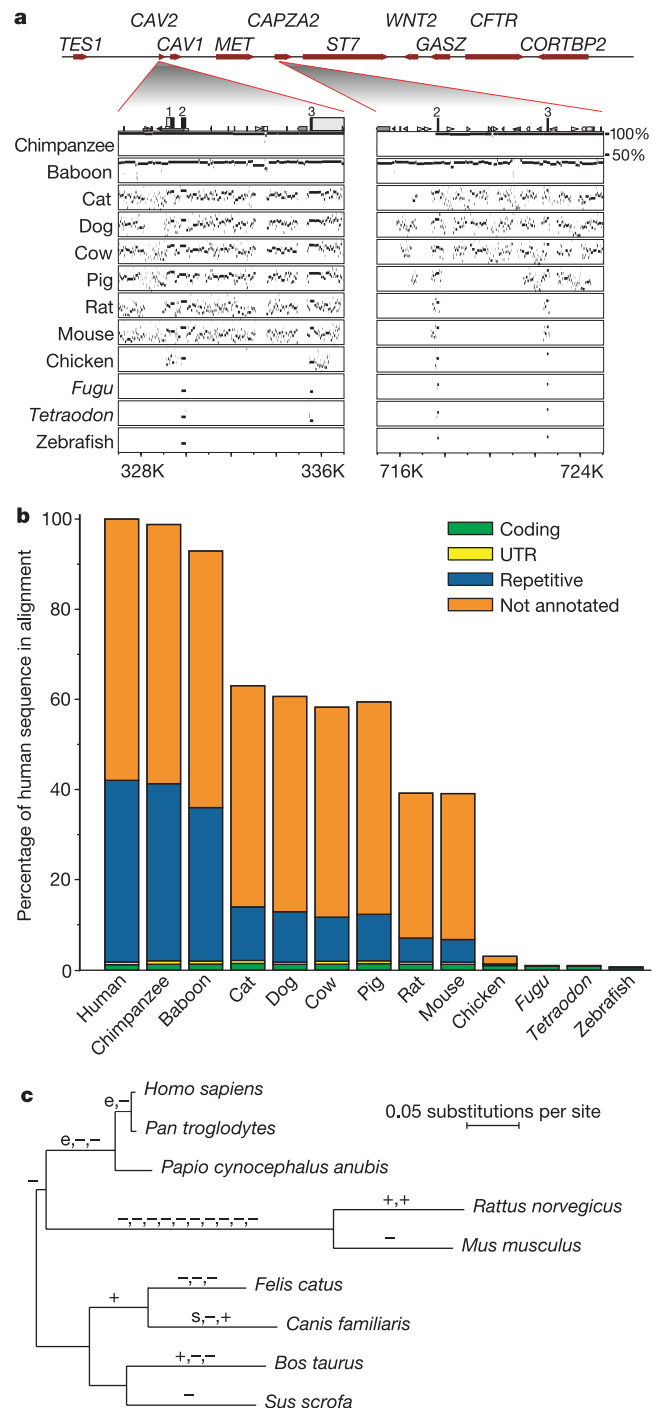


Figure 1 Patterns of sequence conservation. **a**, MultiPipMaker results for the greater *CFTR* region. Pair-wise alignments between the human reference sequence and the sequence of each indicated species were generated. The percentage identity of each gap-free alignment is indicated. Numbered boxes correspond to exons. **b**, Proportion of aligned human sequence in each of four annotated categories, shown for alignments with the sequence of each indicated species. **c**, Deduced phylogenetic tree of indicated mammalian species (see Supplementary Information). Labels on branches reflect differences in exon lengths, as determined manually by parsimony (+, insertion; -, deletion; e, extension due to alteration of splice site or stop codon; s, early stop codon).

Table 1 Sequence generated for the greater *CFTR* region in 12 non-human vertebrates

Species	Non-redundant sequence (total bp)	Clone gaps*	Sequence gaps†
Chimpanzee	1,317,858	2	0
Baboon	1,508,413	0	2
Cat	1,357,338	2	3
Dog	1,195,669	4	10
Cow	1,480,745	0	0
Pig	1,077,879	1	0
Rat	1,600,751	0	3
Mouse‡	1,486,509	0	0
Chicken	415,528	0	0
Fugu§	273,624	0	0
Tetraodon	257,833	0	0
Zebrafish	162,514	1	0

*Number of gaps between non-overlapping BACs within the clone tiling path.
†Number of gaps within the sequenced BACs.
‡Includes ~150 kb of sequence that we generated previously (GenBank accession number AF162137).
§Includes ~36 kb of sequence generated by others (GenBank accession number AJ009961.1).
||Includes some non-orthologous sequence (see Supplementary Information for additional details).

active at around the time of the eutherian radiation, on the basis of the reconstructed phylogenies and divergence levels of the elements themselves. We found no insertions supporting alternative phylogenies (for example, rodents as an outgroup to primates, artiodactyls and carnivores). For example, out of 81 identified segments present in the primates, artiodactyls and carnivores but missing from rodents, none is an identifiable transposable element; instead, all seem to be deletions in the rodent lineage. These analyses seem to definitively refute alternative phylogenies that place the rodents as an outgroup to the other mammals studied here^{16,17}.

Tabulation of exon-length differences indicate a significant excess of deletions relative to insertions, which is particularly strong in the rodent lineage (consistent with previous reports for mouse genes⁶). When this excess is taken into account, the most parsimonious interpretation of the differences again favours the grouping of primates with rodents (Fig. 1c).

Neutral substitution rates estimated from sites in ancestral repeats, which are relics of transposons inserted before the eutherian radiation, also show a higher rate of substitution in the rodent

lineage (Fig. 1c). The branch lengths that we estimate from ancestral repeat sites total about 1.2 substitutions per site in the mammals and are similar to previous calculations (made with these sequence data but with a different multi-sequence alignment¹⁸) that were based on examining untranslated regions (UTRs; total 0.93 substitutions per site), non-exonic regions (total 1.35 substitutions per site) and synonymous substitutions in codons (total 1.28 substitutions per codon).

It is important to note, however, that there are several currently unresolved methodological issues regarding substitution analysis of non-coding genomic sequences. Alignments involving the more diverged sequences tend to have significant uncertainties, and current substitution models do not easily accommodate several known complexities in neutral mutational patterns, including context effects¹⁹, positional variation in substitution rates^{6,20} and the fact (discovered with these sequence data²¹) that there are substitution rate asymmetries associated with transcribed regions. As a result, these rate estimates must be interpreted with caution. Together, these data show that combined molecular evolution studies examining transposon events, neutral substitutions and exonic changes provide a more robust and informative phylogenetic analysis than any method alone.

Of special interest are small genomic regions that are more highly conserved across multiple species than are neutrally evolving sequences, as these may be under purifying selection (that is, selection against mutation of the base) for a functional role. By the use of two methods (E.H.M. *et al.*, manuscript in preparation, and Supplementary Information), we identified 'multi-species conserved sequences' (MCSs) across the greater *CFTR* region. Each method was calibrated such that 5% of bases in the region fell within an MCS (a value chosen to be consistent with the estimated 5% of the mammalian genome under selection based on human–mouse sequence comparisons⁶). Nearly 80% of the MCSs identified with the two methods overlap, and 75% of the MCS bases are identical.

We examined the 1,194 MCSs that overlapped between the two methods. These average 58 base pairs (bp) and represent 3.7% of the bases in the region. About 2% of MCS bases fall in ancestral repeats (corresponding to ~0.4% of the ancestral repeat sequence in the region), suggesting that the fraction of neutrally evolving sequence falsely identified as conserved is small. A total of 32% of the MCS bases overlap known coding sequence or UTRs, involving 98% (125/128) of the known coding exons and 67% (14/21; note that the 5' UTR of the *MET* gene spans two exons) of the known UTRs. The MCSs contain 90.4% and 27.2% of coding and UTR bases, respectively. The relative paucity of UTR bases in MCSs might reflect the fact that these regions include unselected bases or are under lower selection or occasional positive selection, or a combination of these. The remaining 68% of the MCS bases are outside known exons, and virtually all of these (92%) are in MCSs that contain less than 5% repetitive sequence and are present in a single-copy fashion in the human genome (see Supplementary Information). Of the non-exonic MCSs, 16 out of 966 (1.7%; averaging 31 bp) fall within the 1-kilobase (kb) segments immediately upstream of a known transcription start site (the presumed location of most core promoters), which is about 2.3 times more frequent than would be expected if the MCS bases were randomly distributed.

Interestingly, 950 of the 1,194 MCSs (80%; averaging 44 bp) are neither exonic nor lie less than 1-kb upstream of transcribed sequence. Of these, 648 fall in introns and 302 are intergenic; this represents a 28% enrichment of MCS bases in introns (as compared with a random distribution). The detected MCSs overlap with 63% of the functionally validated regulatory elements in the region and 26% of promoters predicted by *in silico* analyses (see Supplementary Information). Several factors could account for the failure of MCSs to overlap all such elements: many of these elements are notably small (<14 bp), whereas our methods do not identify MCSs of less

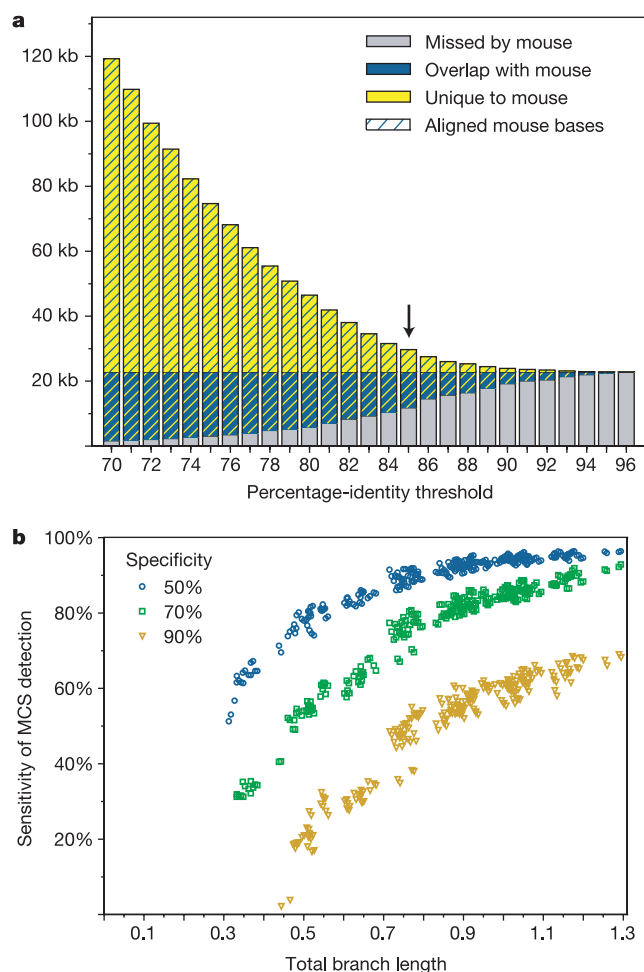


Figure 2 Detection of MCSs by using different mammalian sequences. **a**, For each indicated percentage-identity threshold (see Supplementary Information), the number of aligned mouse bases that overlap a set of 598 MCSs (see text) is shown in blue. Also shown are the number of aligned mouse bases that do not overlap MCSs (yellow) and the number of MCS bases that do not overlap aligned mouse sequence (grey). The arrow indicates the percentage-identity threshold that results in a sensitivity of 48% (with a specificity of 62%; see text). **b**, The relationship between the total branch length of the phylogenetic tree relating the species and the sensitivity of MCS detection is shown for all possible human-containing subsets of the nine mammalian sequences (see Supplementary Information). Results are plotted for three different specificities.

than 25 bp; some of the regulatory elements may be specific to the primate lineage; and the presence or position of some of the annotated elements may be incorrect. Note that most (98%) non-exonic MCSs do not correspond to currently known regulatory elements, yielding a rich supply of candidates for future functional studies.

An important issue relevant to future genome sequencing projects is the degree to which the detection of highly conserved sequences depends on the particular set of species being studied. In the absence of a comprehensive catalogue of functional elements for any large region of the human genome, we used the detection of MCSs as a surrogate for the ability of a species' sequence to identify functionally important regions. Because a draft mouse genome sequence is now available⁶, we first examined the ability of human–mouse pair-wise alignments to detect a set of 561 MCSs in a portion of the greater *CFTR* region for which sequence coverage in all species was nearly complete (see Supplementary Information). Adjustments to the stringency of the human–mouse alignments were ineffective at accurately identifying these MCSs (Fig. 2a). For example, at a percentage-identity threshold of 85%, the sensitivity (percentage of MCS bases that overlap aligned mouse sequence) is only 41%, whereas the specificity (percentage of the aligned mouse sequence that overlaps MCSs) is 77%. This is consistent with the observation that individual conserved elements often cannot be reliably detected with only the human and mouse sequences⁶.

To explore more broadly how MCS detection is dependent on the specific sequences used in the analysis, we identified MCSs with each possible subset of species (always including human) and in each case calibrated the results to yield a defined specificity (percentage of bases overlapping the above 561 MCSs; see Supplementary Information). For the various subsets of mammalian species, there is strong correlation between total divergence of the subset (as measured by the combined branch length of the phylogenetic tree defined by the specific subset of species) and the ability to detect MCSs (Fig. 2b). The results at 90% specificity show that eliminating chimpanzee and baboon does not reduce the number of detected MCS bases. Eliminating the non-human primates, chicken and fish—thereby retaining only the six non-primate mammals in addition to human—reduces the number by 17%. With only one mammal from each major lineage (baboon, cat, cow and mouse), the number is reduced by 29%. Of note, chicken sequence alone detects 40% of MCS bases (representing 94% of the coding but only 29% of the non-coding MCS bases), and this value is higher than for any other single species. Fish sequences alone are effective at detecting coding exons but miss most non-coding MCSs.

The trends in Fig. 2b suggest that most MCSs are broadly distributed among the mammalian lineages, because the power to detect them seems to depend mainly on the total divergence of the subset of species rather than on the particular distribution of the species among lineages. In addition, the fact that at high specificity the sensitivity increases steadily throughout the full range of branch lengths suggests that we may be far from saturating the ability to detect such conserved sequences, even with the full set of mammals examined here. It thus seems that combined branch length will be a useful metric for guiding the selection of additional genomes to sequence.

The nature of the mutational events that have produced the observed differences in the greater *CFTR* region among the sequenced species is of fundamental evolutionary interest. Despite strict conservation in gene order and content and the absence of any observed syntenic breaks, there is significant variation in the amount of non-coding sequence, suggesting variation in genome expansion or compression. For example, the region is about tenfold smaller in the two pufferfish species than in the mammals, which themselves vary by as much as ~15% (see Supplementary Information). These findings are consistent with the relative genome sizes established for some of these species^{3,5,6} and point to significant

changes in this genomic region throughout vertebrate evolution. To account for these differences, we looked for evidence of molecular events that have contributed to genome expansion or compression.

The fraction of sequence corresponding to interspersed repeats, which are remnants of insertion events mediated by active or extinct transposons, varies from roughly 29% to 39% among the mammals (Fig. 3a; see Supplementary Information). The interspersed repeat content is much lower in the non-mammalian vertebrates, with the pufferfish and zebrafish sequences containing less than 1% and 13% interspersed repeats, respectively. In addition, the distribution of interspersed repeat types differs greatly among species (Fig. 3a).

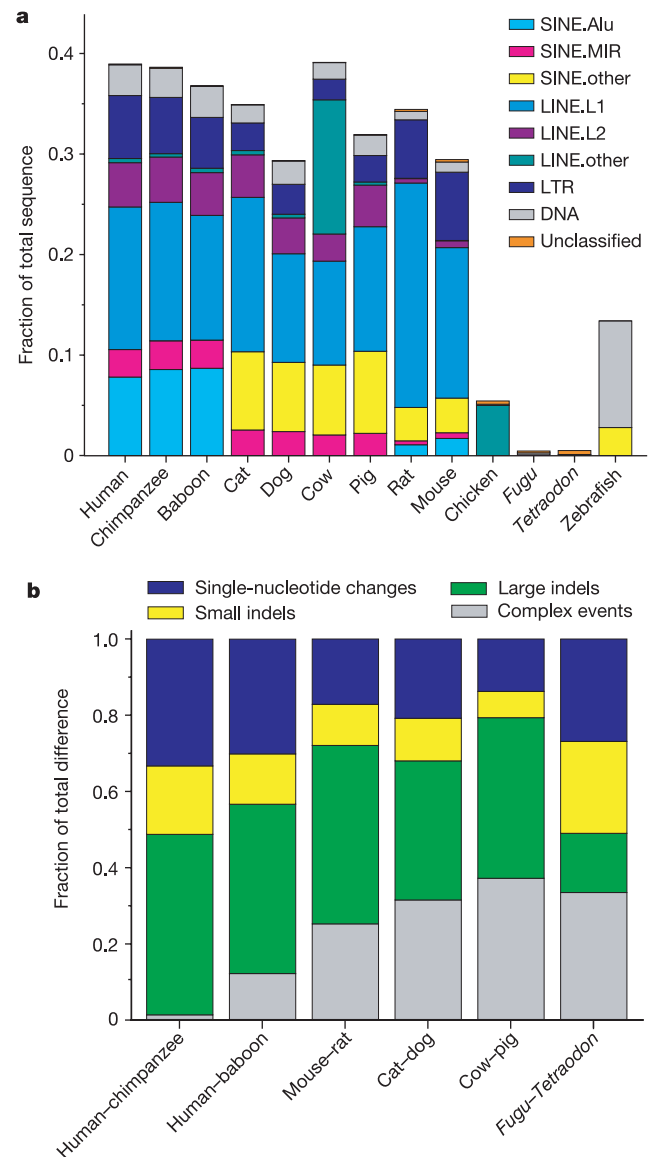


Figure 3 Comparison of genome dynamics among species. **a**, Relative content of different interspersed repeat types plotted for each species. **b**, Relative contribution of different mutational events within major lineages. Sequence alignments between the indicated species pairs were used to classify all sequence differences: single-nucleotide changes, small indels (<100 bp), large indels (>100 bp) and complex events (involving >100 bp and most probably resulting from an inversion or more than one deletion and/or insertion; see Supplementary Information). The relative fraction of nucleotide differences falling into each class is shown for the indicated species pairs. The overall percentage of single-nucleotide mismatches for the greater *CFTR* region in each species pair is 1.15%, human–chimpanzee; 5.60%, human–baboon; 13.93%, mouse–rat; 15.57%, cat–dog; 17.52%, cow–pig; and 20.50%, *Fugu*–*Tetraodon*. SINE, short interspersed nuclear element; LINE, long interspersed nuclear element.

Thus, the accumulation of interspersed repeats correlates with expansion of the greater *CFTR* region in the main lineages.

Differences in the pattern of interspersed repeat types are also apparent between and within mammalian lineages. Indeed, the accumulation of species-specific interspersed repeats seemingly accounts for much of the differences in relative size and interspersed repeat composition of this genomic region within these lineages (see Supplementary Information). Our findings implicate an increased rate of interspersed repeat insertions rather than variation in the extent of deletions as a primary cause of the observed size differences for this region; this contrasts with the observed higher rate of deletion versus insertion in the mouse lineage, which has resulted in a significantly smaller size of the mouse genome relative to human⁶.

For the three primate species, we estimated the rate of large (>100 bp) insertions or deletions (indels) using parsimony (see Supplementary Information). This revealed variable rates of insertion and deletion, producing a slight relative expansion in the human lineage and a slight compression in both the chimpanzee and baboon lineages. We also examined the relative importance of interspersed repeat insertions and large indels compared with small indels (<100 bp) and single-nucleotide changes. Although most mutational events leading to human–chimpanzee differences are single-nucleotide changes, they account for only 33% of the bases that differ between these two species (Fig. 3b and Supplementary Information). Indeed, nearly half of the differing bases correspond to large indels (Fig. 3b; as an example, note the large deletion in the chimpanzee sequence immediately upstream of *CAPZA2* exon 2 in Fig. 1a). Similarly, at least 44% of the bases that differ between the human and baboon sequences reflect large indels. Thus, among the primates, large indels are the principal mechanism accounting for the observed sequence differences, a finding that is consistent with other studies^{22,23}.

Similar analyses of the other mammals identified a similar spectrum of mutational events, with large indels accounting for the largest fraction of bases that differ in each lineage (Fig. 3b). This suggests that the non-aligning regions in mammalian sequence primarily reflect the insertion of repetitive elements and the deletion of ancestral sequences. Given this assumption, the alignment of the human greater *CFTR* region to the other mammalian sequences suggests that a minimum of 63% of this segment of the human genome was present in the last common ancestor of the mammals sampled here, some 94 million years ago²⁴, and that the rodents represent the most derived mammalian sequence (minimum 39% ancestral; Fig. 1b). Finally, the pufferfish sequences show a distinct pattern of genomic changes (Fig. 3b). As compared with the mammals, single-nucleotide mismatches and small indels are more prominent than large indels. In part, this probably reflects the paucity of transposon insertions in pufferfish and the deleterious consequences of large indels in a more compact genome.

In summary, our approach for multi-species sequence generation and analysis is providing a previously unavailable glimpse through the window of vertebrate evolution. Our findings confirm that the genomes of rodents are changing faster than those of primates, carnivores and artiodactyls, in agreement with many reports^{6,25} but in contrast to studies defending the molecular clock²⁶. In addition, we have found a substantial number of sequence elements conserved across multiple species, most of which are located in non-coding regions. Although the general mutational spectrum is similar among all vertebrate lineages, differences in the relative contribution of the various types of molecular events have sculpted the genome of each species in a unique fashion. It should be noted that the findings reported here pertain to a single genomic region, whereas the vertebrate genome is known to show highly non-uniform patterns of sequence conservation^{6,20}. Our efforts to sequence many other genomic targets in multiple species (see the NISC Comparative Sequencing Program website: <http://www.nisc.nih.gov>) should provide a broader and more informed

view of such regional variation. Together, our studies point to myriad avenues for investigating the evolutionary and functional features of the vertebrate genome. □

Methods

Sequence generation and analysis

Targeted genomic regions of interest were isolated in overlapping bacterial artificial chromosome (BAC) clones²⁷ by using libraries derived from different vertebrate species. We used 'universal' hybridization probes, designed from small stretches of sequence conserved between human and mouse²⁸, to isolate BACs from multiple mammals in parallel. For isolating BACs from non-mammalian vertebrates, species-specific probes (typically designed from available gene sequences) were mostly used for clone isolation. For each species, a minimally overlapping tiling path of BACs spanning the genomic target was selected, and individual BACs were sequenced by a conventional shotgun sequencing strategy (see Supplementary Information).

To facilitate computational analyses, we compiled a single, non-redundant nucleotide sequence for each species by merging together the data generated for all sequenced BACs. This was then annotated to indicate the locations of known genes and coding regions on the basis of matches to human and/or mouse reference cDNA sequences, exons predicted by Genscan²⁹, and repetitive elements using RepeatMasker. Annotated sequence records are available at the NISC Comparative Sequencing Program website (<http://www.nisc.nih.gov/data>). A more detailed description of this assembly and annotation process is given in the Supplementary Information.

The assembled and annotated sequences were then subjected to a series of analyses, starting with the generation of multi-sequence alignments and including the studies described above and in Figs 1–3 (such as establishing orthology, examining the general patterns of sequence conservation, performing phylogenetic analyses, detecting highly conserved sequences and investigating genome dynamics). Details of these analyses are given in the Supplementary Information.

Visualization and dissemination of results

The establishment of robust and user-friendly computational-based approaches for capturing, visualizing and disseminating multi-species sequences and the data emanating from their comparative analyses represents an increasingly important challenge. By using our data as a model, we developed a viable solution to this problem by incorporating our data into the University of California Santa Cruz Genome Browser³⁰. The resulting web-based resource (<http://genome.ucsc.edu>) serves as an additional electronic supplement to this paper; a low-resolution overview of this website is given in the Supplementary Information.

Integrating our data into the UCSC Genome Browser allows it to be visualized within the context of the growing body of information about the human and other genome sequences represented on this browser. We have configured the browser to display custom tracks showing the results of various analyses, including those involving multiple sequences (see Supplementary Information). The dynamic nature of the browser allows convenient navigation from a detected region of conservation to the underlying sequence alignment, as well as the ability to examine the results of comparative analyses using different species' sequence as the reference.

Availability of data and analysis tools

Details about the data underlying the studies reported here (including updated BAC contig maps, information about each sequenced clone, compiled and annotated sequence files for each species and links to the various electronic resources) are available on the NISC Comparative Sequencing Program website (<http://www.nisc.nih.gov/data>). The blastz program and MultiPipMaker network services can be obtained on the Penn State Bioinformatics Group website (<http://bio.cse.psu.edu>).

Received 11 April; accepted 16 June 2003; doi:10.1038/nature01858.

1. Batzoglou, S., Pachter, L., Mesirov, J. P., Berger, B. & Lander, E. S. Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res.* **10**, 950–958 (2000).
2. Roest Crollius, H. et al. Estimate of human gene number provided by genome-wide analysis using *Tetraodon nigroviridis* DNA sequence. *Nature Genet.* **25**, 235–238 (2000).
3. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
4. Chen, R., Bouck, J. B., Weinstock, G. M. & Gibbs, R. A. Comparing vertebrate whole-genome shotgun reads to the human genome. *Genome Res.* **11**, 1807–1816 (2001).
5. Aparicio, S. et al. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297**, 1301–1310 (2002).
6. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
7. Dubchak, I. et al. Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Res.* **10**, 1304–1306 (2000).
8. Gottgens, B. et al. Analysis of vertebrate *SCL* loci identifies conserved enhancers. *Nature Biotechnol.* **18**, 181–186 (2000).
9. Hardison, R. C. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.* **16**, 369–372 (2000).
10. Pennacchio, L. A. & Rubin, E. M. Genomic strategies to identify mammalian regulatory sequences. *Nature Rev. Genet.* **2**, 100–109 (2001).
11. Rommens, J. M. et al. Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science* **245**, 1059–1065 (1989).
12. Felsenfeld, A., Peterson, J., Schloss, J. & Guyer, M. Assessing the quality of the DNA sequence from The Human Genome Project. *Genome Res.* **9**, 1–4 (1999).

13. Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res* **13**, 103–107 (2003).
14. Schwartz, S. *et al.* MultiPipMaker and supporting tools: alignments and analysis of multiple genomic DNA sequences. *Nucleic Acids Res.* **31**, 3518–3524 (2003).
15. Murphy, W. J. *et al.* Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* **294**, 2348–2351 (2001).
16. Poux, C., Van Rheede, T., Madsen, O. & de Jong, W. W. Sequence gaps join mice and men: phylogenetic evidence from deletions in two proteins. *Mol. Biol. Evol.* **19**, 2035–2037 (2002).
17. Huelsenbeck, J. P., Larget, B. & Swofford, D. A compound Poisson process for relaxing the molecular clock. *Genetics* **154**, 1879–1892 (2000).
18. Cooper, G. M. *et al.* Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Res.* **13**, 813–820 (2003).
19. Siepel, A. & Haussler, D. *Proc. 7th Annual Int. Conf. Research in Computational Molecular Biology* (ACM, New York, 2003).
20. Hardison, R. C. *et al.* Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res.* **13**, 13–26 (2003).
21. Green, P. *et al.* Transcription-associated mutational asymmetry in mammalian evolution. *Nature Genet.* **33**, 514–517 (2003).
22. Frazer, K. A. *et al.* Genomic DNA insertions and deletions occur frequently between humans and nonhuman primates. *Genome Res.* **13**, 341–346 (2003).
23. Britten, R. J. Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. *Proc. Natl Acad. Sci. USA* **99**, 13633–13635 (2002).
24. Springer, M. S., Murphy, W. J., Eizirik, E. & O'Brien, S. J. Placental mammal diversification and the Cretaceous/Tertiary boundary. *Proc. Natl Acad. Sci. USA* **100**, 1056–1061 (2003).
25. Li, W. H., Ellsworth, D. L., Krushkal, J., Chang, B. H. & Hewett-Emmett, D. Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. *Mol. Phylogenet. Evol.* **5**, 182–187 (1996).
26. Kumar, S. & Subramanian, S. Mutation rates in mammalian genomes. *Proc. Natl Acad. Sci. USA* **99**, 803–808 (2002).
27. Shizuya, H. *et al.* Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc. Natl Acad. Sci. USA* **89**, 8794–8797 (1992).
28. Thomas, J. W. *et al.* Parallel construction of orthologous sequence-ready clone contig maps in multiple species. *Genome Res.* **12**, 1277–1285 (2002).
29. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
30. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank J. Weissenbach and H. Roest Crolius for *Tetraodon* BACs; M. Diekhans for computational expertise; N. Goldman and Z. Yang for advice on phylogenetic analyses; and F. Collins and J. Mullikin for critically reading the manuscript. We acknowledge the support of the National Human Genome Research Institute (National Institutes of Health) and the Howard Hughes Medical Institute.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E.D.G. (egreen@nhgri.nih.gov). GenBank accession numbers for BAC-derived sequences are provided in the Supplementary Information.

CYLD is a deubiquitinating enzyme that negatively regulates NF- κ B activation by TNFR family members

Eirini Trompouki¹, Eudoxia Hatzivassiliou^{1*}, Theodore Tsichritzis^{1*}, Hannah Farmer², Alan Ashworth² & George Mosialos¹

¹Institute of Immunology, Biomedical Sciences Research Center 'Alexander Fleming', 34 Alexander Fleming Street, Vari 16672, Greece

²Cancer Research UK Gene Function & Regulation Group, The Breakthrough Breast Cancer Research Centre, Institute of Cancer Research, Chester Beatty Laboratories, 237 Fulham Road, London SW3 6JB, UK

* These authors contributed equally to this work

Familial cylindromatosis is an autosomal dominant predisposition to tumours of skin appendages called cylindromas. Familial cylindromatosis is caused by mutations in a gene encoding the CYLD protein of previously unknown function¹. Here we show that CYLD is a deubiquitinating enzyme that negatively regulates activation of the transcription factor NF- κ B by specific tumour-

necrosis factor receptors (TNFRs). Loss of the deubiquitinating activity of CYLD correlates with tumorigenesis. CYLD inhibits activation of NF- κ B by the TNFR family members CD40, XEDAR and EDAR in a manner that depends on the deubiquitinating activity of CYLD. Downregulation of CYLD by RNA-mediated interference augments both basal and CD40-mediated activation of NF- κ B. The inhibition of NF- κ B activation by CYLD is mediated, at least in part, by the deubiquitination and inactivation of TNFR-associated factor 2 (TRAF2) and, to a lesser extent, TRAF6. These results indicate that CYLD is a negative regulator of the cytokine-mediated activation of NF- κ B that is required for appropriate cellular homeostasis of skin appendages.

Activation of NF- κ B by cytokines depends on the activity of the I κ B kinase (IKK) complex, which mediates phosphorylation of the NF- κ B inhibitor I κ B². NEMO (also known as IKK- γ) is an essential component of the IKK complex. IKK-mediated phosphorylation of I κ B causes its degradation by the ubiquitin-proteasome pathway and results in nuclear translocation and activation of NF- κ B.

A yeast two-hybrid screen with NEMO as the bait identified complementary DNAs encoding either the alternatively spliced, 'full-length' CYLD of 953 amino acids or two truncated forms of the protein spanning amino acids 211–709 and 387–953. The interaction of CYLD with NEMO was specific, because full-length CYLD did not interact with two mutated forms of NEMO (C417R and D406V) that have an altered carboxy-terminal zinc-finger motif³. To determine whether this association occurs in mammalian cells, Flag-tagged CYLD and glutathione S-transferase (GST)-tagged NEMO were coexpressed in human embryonic kidney (HEK) 293T cells. Co-precipitation experiments indicated that CYLD and NEMO can indeed interact in mammalian cells after their coexpression (Fig. 1a, b). The amino-terminal region of CYLD mediates the interaction with NEMO, because deletion of the N-terminal 537 amino acids of CYLD abolished NEMO binding (Fig. 1c). We could not detect an interaction between endogenous CYLD and NEMO, either because of the possible transient nature of this interaction or because of the relatively low affinity of our antisera against CYLD.

The C-terminal 365 amino acids of CYLD contain motifs found in ubiquitin-specific proteases (UBPs), a subclass of deubiquitinating enzymes thought to be responsible for the removal of polyubiquitin chains from polypeptides^{1,4}. Most pathogenic inactivating mutations of CYLD result in truncations or frameshift alterations of the C-terminal region of the molecule¹. To determine whether CYLD can act as a UB, Flag-tagged wild-type CYLD and three C-terminally truncated mutants of CYLD that are associated with familial cylindromatosis (encoding residues 1–932, 1–864 and 1–754 of CYLD; see Online Mendelian Inheritance in Man number 132700 at <http://ncbi.nlm.gov/Omim/>) were expressed in HEK 293T cells, immunoprecipitated and tested for their ability to cleave tetraubiquitin. Wild-type CYLD readily cleaved the tetrameric substrate to its monomer, dimer and trimer; however, the pathogenic mutations associated with familial cylindromatosis abolished the UB activity of CYLD, indicating a correlation between tumorigenesis and loss of the deubiquitinating activity of CYLD (Fig. 1d, top). The deubiquitinating activity of CYLD did not require any other mammalian proteins, because it was detectable in bacteria (Supplementary Fig. 1). In addition, the deubiquitinating activity of CYLD was abolished by replacing the conserved catalytic residue Cys 601 with serine.

The interaction of CYLD with NEMO prompted us to investigate the potential involvement of CYLD in NF- κ B activation. CD40, XEDAR and EDAR are members of the TNFR family implicated in the proper development of skin appendages from which cylindromas arise^{3,5,6}. Expression of wild-type CYLD in HEK 293T cells inhibited the ability of coexpressed CD40 (Fig. 2a), XEDAR or EDAR (Supplementary Fig. 2) to activate the NF- κ B pathway; by

contrast, the CYLD mutant associated with familial cylindromatosis that lacks 21 amino acids at its C terminus was markedly impaired in its ability to inhibit NF- κ B activation by these receptors. CYLD(538–953), which is catalytically active but does not bind to NEMO, also failed to inhibit CD40-mediated activation of NF- κ B, suggesting that the NF- κ B inhibitory activity of CYLD is dependent on NEMO (Fig. 2b). In addition, wild-type CYLD inhibited activation of NF- κ B induced by the Epstein-Barr virus oncoprotein LMP1, which mimics activated CD40 (ref. 7 and data not shown).

To test whether endogenous CYLD functions as an inhibitory factor of NF- κ B, we used a short interfering RNA (siRNA) approach to reduce the expression of CYLD and determined the effects on the basal and CD40-ligand-induced activity of NF- κ B. HEK 293T cells were transfected with an NF- κ B-dependent luciferase reporter plasmid and either a human CYLD-specific siRNA (HCYLDsiRNA) or an unrelated siRNA (NSsiRNA). Transfection of HCYLDsiRNA resulted in a marked decrease both in endogenous CYLD messenger RNA (Supplementary Fig. 3a) and in the amount of overexpressed Flag-tagged CYLD protein (Supplementary Fig. 3b). Both the basal and the CD40-ligand-induced activity of NF- κ B were significantly augmented on reduction of CYLD expression, reinforcing the notion that CYLD acts as a negative regulator of the NF- κ B pathway (Fig. 2c).

To analyse the mechanism underlying the inhibition of TNFR-

mediated NF- κ B activation by CYLD, we studied the point in the signalling pathway at which CYLD interferes. TRAF2 and TRAF6 have been implicated in the activation of NF- κ B by CD40, EDAR, XEDAR and LMP1 (refs 6, 8–13). TRAF2 and TRAF6 interact directly or indirectly with the cytoplasmic tail of the receptor. These proteins are activated by oligomerization, and this event can be induced either by their overexpression or by receptor activation. After receptor activation, the principal rate-limiting step of NF- κ B induction is activation of the IKK complex, which can be also induced by overexpressing either catalytic subunit of IKK.

Wild-type CYLD inhibited TRAF2-mediated NF- κ B activation markedly (by about 93%; Fig. 3a); however, the deubiquitinase-deficient mutants of CYLD (CYLD(1–932) and CYLD-C601S) were severely impaired in this activity (Fig. 3a and Supplementary Fig. 4). Wild-type CYLD had a weaker but still significant inhibitory effect (about 50% inhibition) on NF- κ B activation by TRAF6, as compared with its effect on activation by TRAF2 (Fig. 3b). Notably, CYLD had no effect on IKK β -mediated activation of NF- κ B (Fig. 3c). These results indicate that the principal target of CYLD is located upstream of IKK β . These findings are also consistent with

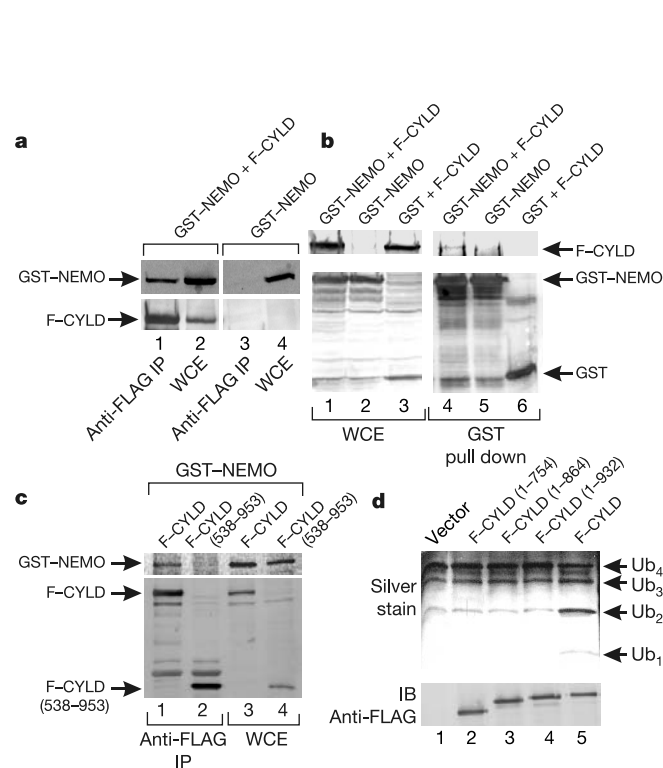


Figure 1 CYLD is a deubiquitinating enzyme that interacts with NEMO. **a–c**, HEK 293T cells were transfected with vectors expressing GST–NEMO or GST in the presence or absence of constructs expressing Flag-tagged (F–) CYLD or CYLD(538–953). Cell lysates were precleared and subjected to either immunoprecipitation with an antibody against Flag (anti-Flag IP) or a GST pull-down assay. Along with an aliquot of the whole-cell extract (WCE), the samples were subjected to immunoblotting with antisera against GST (**a** and **c**, top), the M5 monoclonal antibody against Flag (**a** and **c**, bottom) or antisera against GST and against CYLD (**b**). **d**, Flag-tagged full-length CYLD or truncations associated with familial cylindromatosis were immunoprecipitated with the M2 antibody against Flag from lysates of transfected HEK 293T cells and tested for their ability to cleave tetraubiquitin. The reaction mixture was analysed by tricine SDS–PAGE and silver staining (top). Arrows indicate the positions of tetrameric (Ub₄), trimeric (Ub₃), dimeric (Ub₂) and monomeric (Ub₁) ubiquitin. The amount of immunoprecipitated Flag-tagged proteins was determined by immunoblotting with the M5 antibody (bottom).

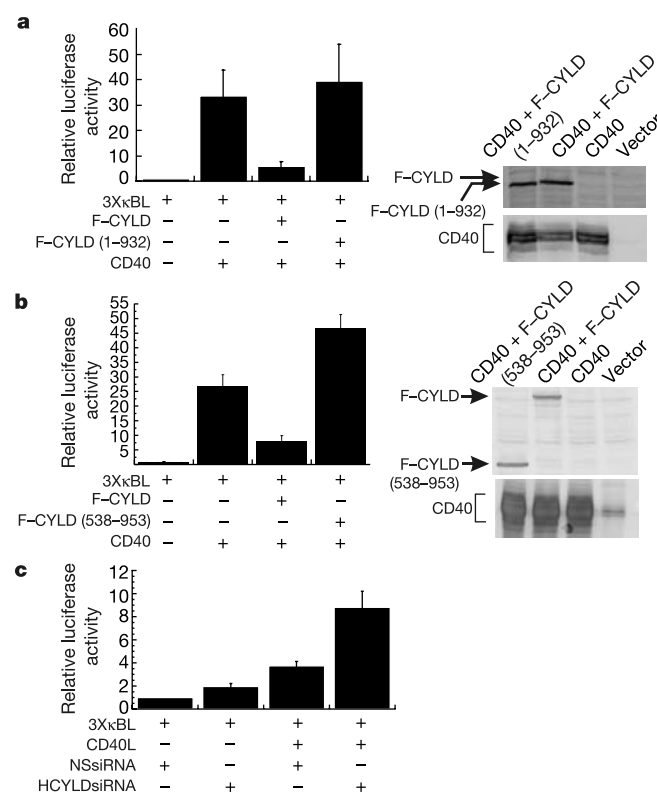


Figure 2 Effect of CYLD and truncated forms of CYLD on NF- κ B activation by CD40. HEK 293T cells were transfected with the 3X κ BL and the pGK– β -galactosidase reporter plasmids plus a CD40 expression vector in the presence or absence of vectors expressing Flag-tagged (F–) CYLD or CYLD(1–932) (**a**); a CD40 expression vector in the presence or absence of vectors expressing Flag-tagged CYLD or CYLD(538–953) (**b**); or the same amount of either a nonspecific siRNA (NSsiRNA) or a human CYLD-specific siRNA (HCYLDsiRNA) (**c**). Cell extracts were prepared either 24 h (**a**, **b**) or 48 h (**c**) after transfection, and their luciferase and β -galactosidase activities were determined. In **c**, the cells were left untreated or stimulated with CD40 ligand for 5 h before lysis. Results are the mean $\pm$ s.d. relative luciferase activity from three (**a**, **b**) or four (**c**) independent experiments. Representative immunoblots of whole-cell lysates from equal numbers of HEK 293T cells subjected to transfection in one representative experiment are shown on the right in **a** and **b**.

our observed cytoplasmic localization of CYLD tagged with either green fluorescent protein (GFP) or the Flag epitope in epithelial cells (data not shown).

The NF- κ B-inducing activity of TRAF2 and TRAF6 depends on their RING-finger-dependent ubiquitin ligase activity and auto-ubiquitination^{14,15}. The autoubiquitination of TRAF2 and TRAF6 depends on the activity of the ubiquitin-conjugating enzyme Ubc13. This enzyme mediates the formation of polyubiquitin chains linked through Lys 63, which do not mediate degradation of the polyubiquitinated protein^{14–17}. A dominant-negative mutant of Ubc13 inhibited activation of NF- κ B by CD40, XEDAR and EDAR, consistent with the idea that both Ubc13 and polyubiquitination through Lys 63 have a crucial role in the activation of NF- κ B by these receptors (Supplementary Fig. 5a, b).

A possible mechanism for the inhibitory activity of CYLD on TNFR and TRAF-mediated NF- κ B activation might be based on the CYLD-dependent deubiquitination of TRAF2 and/or TRAF6. We tested this idea by determining the extent of TRAF2 and TRAF6 autoubiquitination on expression of wild-type or UBP-deficient CYLD. Flag-tagged TRAF2 was coexpressed with haemagglutinin A (HA)-tagged ubiquitin in HEK 293T cells and immunoprecipitated under stringent conditions that would disrupt most protein–protein interactions. Polyubiquitination of TRAF2 was apparent by immunoblot analysis and was nearly abolished by the expression of a dominant-negative mutant of Ubc13, consistent with previous

reports^{14,16} of non-degradative Lys-63-linked polyubiquitination of TRAF2 under similar conditions (Fig. 4a and Supplementary Fig. 5).

Expression of wild-type CYLD reduced TRAF2 polyubiquitination to roughly background levels (Fig. 4a). By contrast, the CYLD(1–932) deubiquitinase-deficient mutant caused a small increase in TRAF2 polyubiquitination, consistent with a possible dominant-negative effect (Fig. 4a). Deubiquitination of TRAF2 by CYLD did not affect the steady-state quantities of protein (data not shown). In addition, the proteasome inhibitor ALLN did not affect the ubiquitination or steady-state amounts of TRAF2 under the conditions of our experiments (data not shown). These findings are consistent with the notion that CYLD targets the Lys-63-linked non-degradative polyubiquitination of TRAF2, which may be related to its ability to activate NF- κ B¹⁴. The effect of CYLD on TRAF2 ubiquitination was specific, because expression of CYLD did not affect the overall extent of ubiquitination in the cell

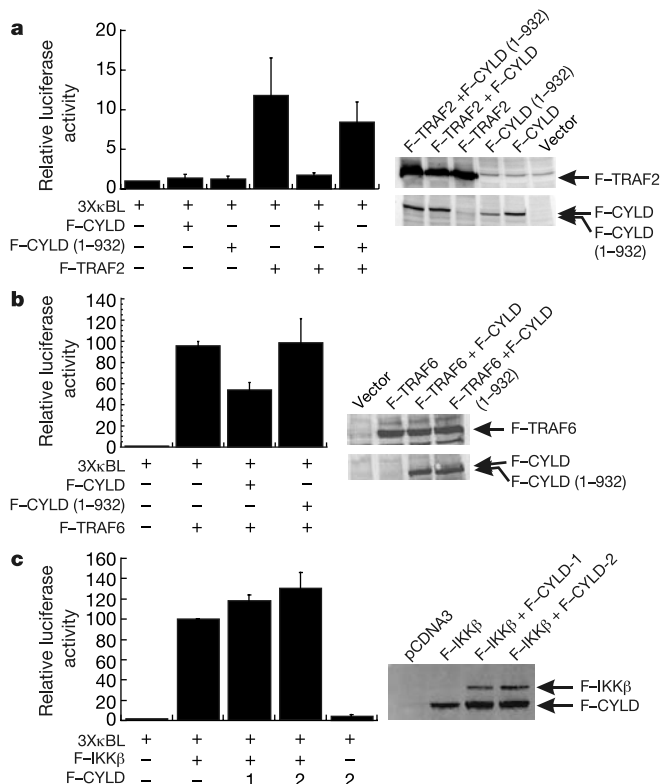


Figure 3 Effect of CYLD and CYLD(1–932) on NF- κ B activation by TRAF2, TRAF6 and IKK β . HEK 293T cells were transfected with the 3XkBL and the pGK- β -galactosidase reporter plasmids plus vectors expressing Flag-tagged (F-) TRAF2, CYLD or CYLD(1–932) (**a**); Flag-tagged TRAF6, CYLD or CYLD(1–932) (**b**); and Flag-tagged IKK β and the indicated amounts (in μ g) of the Flag-tagged CYLD expression vector (**c**). After 24 h, cell extracts were prepared and the luciferase and β -galactosidase activities were determined. Results are the mean $\pm$ s.d. relative luciferase activity from three independent experiments. Representative immunoblots of whole-cell lysates from an equal number of HEK 293T cells subjected to transfection are shown on the right.

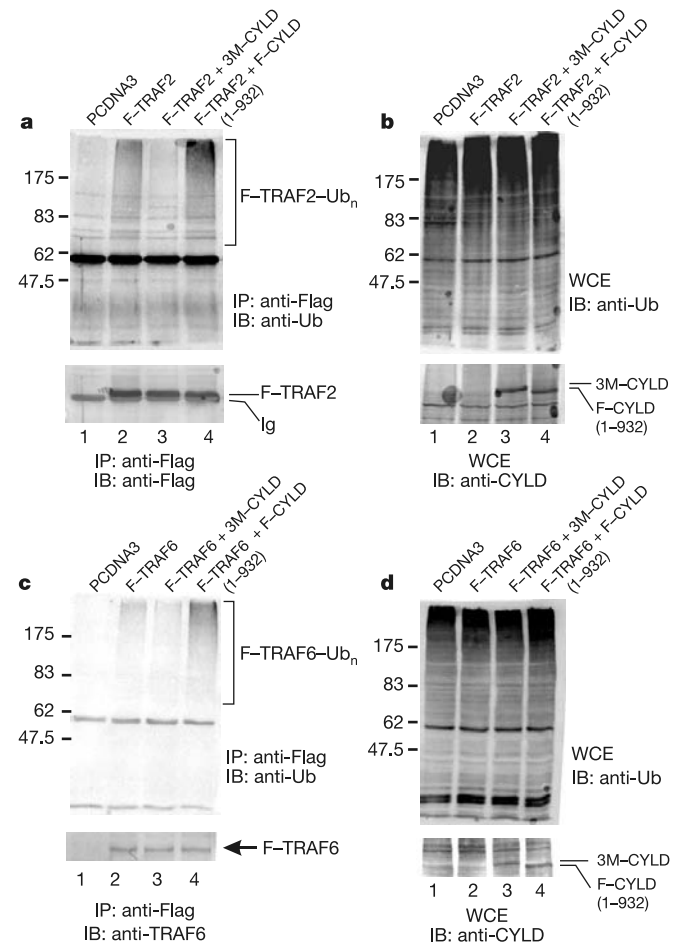


Figure 4 Effect of CYLD and CYLD(1–932) on TRAF2 and TRAF6 polyubiquitination. HEK 293T cells were transfected with expression vectors for HA-tagged ubiquitin, Flag-tagged (F-) TRAF2, TRAF6 and CYLD(1–932), and Myc-tagged CYLD (3M-CYLD). Proteins immunoprecipitated with the M2 antibody against Flag from lysates of transfected cells prepared 24 h after transfection were analysed by immunoblotting with antisera against ubiquitin (**a** and **c**, top) and then reprobbed with the M2 antibody against Flag (**a**, bottom) or an antibody against TRAF6 (**c**, bottom). The positions of polyubiquitinated Flag-tagged TRAF2 (F-TRAF2-Ub_n) and TRAF6 (F-TRAF6-Ub_n), and M2 immunoglobulin (Ig) are shown. Whole-cell extracts (WCE) of the transfected cells were subjected to SDS-PAGE and immunoblotting with a monoclonal antibody against ubiquitin to determine the overall extent of ubiquitination (**b** and **d**, top). The same blots were reprobbed with antisera against CYLD (**b** and **d**, bottom).

(Fig. 4b), nor did it reverse the E6-mediated degradation of p53 by ubiquitination (data not shown). Wild-type CYLD caused a weak but reproducible reduction in TRAF6 polyubiquitination (Fig. 4c), whereas CYLD(1–932) had a small positive effect (Fig. 4c). These experiments suggest that CYLD inhibits the NF- κ B activation pathway, at least in part, by removing Lys-63-linked polyubiquitin from TRAF2 and TRAF6. In addition, CYLD might inhibit the NF- κ B pathway by deubiquitinating I κ B, although we do not have direct evidence to support this mechanism.

Our studies have identified a regulatory mechanism of the NF- κ B pathway that relies on the deubiquitinating activity of CYLD. This protein mediates an inhibitory action on the NF- κ B pathway by reversing the ubiquitination of TRAF2 and TRAF6, possibly through a transient interaction with NEMO. Binding of CYLD to NEMO probably enables CYLD to position itself on IKK holoenzyme that has been translocated onto the cytoplasmic tail of the activated TNFR, thereby allowing CYLD to deubiquitinate receptor-bound TRAF proteins^{18–20}. CYLD might be activated by binding to the ubiquitin moiety of target proteins, as this has been reported for another deubiquitinase²¹. Our results indicate that increased or prolonged activation of NF- κ B caused by a loss of CYLD is important in the generation and growth of cylindromas, possibly because of the inhibition of apoptosis^{22,23}. Because of the localized and accessible nature of these tumours, inhibitors of the NF- κ B pathway such as steroidal and nonsteroidal anti-inflammatory compounds are candidate therapeutic agents for cylindromas. □

Methods

Plasmid construction

Plasmids were obtained or constructed by standard methods and details are given in the Supplementary Information.

Yeast two-hybrid screen

A plasmid expressing a fusion of GAL4 DNA-binding domain to full-length human NEMO (pBridgeNEMO) was used as the bait plasmid. We carried out the yeast two-hybrid screen as described²⁴.

Cell lines, transfection assays and reporter assays

The HEK 293T cell line was maintained in Dulbecco's modified Eagle's medium supplemented with 10% fetal bovine serum (D10, Invitrogen) at 37 °C in 5% CO₂. HEK 293T cells were transfected with ExGEN500 (Fermentas) or Lipofectamine 2000 (Invitrogen) transfection reagents. We determined the activation of NF- κ B by using an NF- κ B-dependent luciferase reporter plasmid 3X κ BL (ref. 25). Normalization for transfection efficiency in luciferase reporter assays was done by including a constant amount of a β -galactosidase expression plasmid (pGK- β -gal)²⁶ in all transfections and measuring β -galactosidase activity. Luciferase and β -galactosidase activities were assayed 1–2 d after transfection. Relative luciferase activities are expressed as fold activation relative to the activity of the NF- κ B-dependent luciferase reporter (3X κ BL) alone, and were calculated by dividing the values for luciferase activity by the corresponding values for β -galactosidase activity. For luciferase-reporter assays, cells were lysed in reporter lysis buffer (Promega) and luciferase activity was assayed with the luciferase assay system (Promega) according to the manufacturer's instructions. We measured β -galactosidase activity with the Galacton-Plus substrate system (Tropix) according to the manufacturer's instructions.

Deubiquitination assays

Wild-type or mutated Flag-tagged CYLD proteins were immunoprecipitated from HEK 293T cell lysates prepared with Nonidet P-40 (NP-40) lysis buffer (25 mM Tris-HCl pH 7.5, 150 mM NaCl, 5 mM EDTA, 10% glycerol, 0.5% NP-40, 1 mM phenyl methylsulphonyl fluoride) from cells that were pretreated for 4 h with 20 μ M ALLN. The immunoprecipitated proteins were incubated with 1.25 μ g of tetraubiquitin (Affinity Research Products) in 15 μ l of 50 mM Tris-HCl buffer (pH 7.2) containing 1 mM dithiothreitol (DTT) for 2 h at 37 °C. Reaction mixtures were analysed by SDS-PAGE, followed by silver staining or immunoblotting with a monoclonal antibody against Flag.

GST pull-down assays, immunoprecipitation and immunoblotting

Cells were lysed for 30 min on ice in 500 μ l of NP-40 lysis buffer for the GST pull-down assay shown in Fig. 1c and the immunoprecipitations shown in Fig. 1b, d, or in 500 μ l or 1,000 μ l of RIPA buffer (NP-40 lysis buffer containing 1 mM DTT, 0.1% SDS and 1% deoxycholic acid) for the other immunoprecipitations. We carried out GST pull-down assays and immunoprecipitations as described²⁷. Immunoblots were developed with the ECL-plus reagent (Amersham Biosciences) according to the manufacturers instructions. Antibodies were obtained from Santa Cruz Biotechnology, with the exception of the M5

and M2 antibodies against Flag, which were obtained from Sigma. We purchased sheep secondary antibodies against mouse and donkey secondary antibodies against rabbit conjugated to horseradish peroxidase from Amersham Pharmacia.

RNA-mediated interference

RNA-mediated interference for downregulating CYLD expression was done by the transfection of double-stranded RNA oligonucleotides, details of which are given in the Supplementary Information.

Received 17 March; accepted 14 May 2003; doi:10.1038/nature01803.

- Bignell, G. R. et al. Identification of the familial cylindromatosis tumour-suppressor gene. *Nature Genet.* **25**, 160–165 (2000).
- Ghosh, S. & Karin, M. Missing pieces in the NF- κ B puzzle. *Cell* **109**, S81–S96 (2002).
- Jain, A. et al. Specific missense mutations in NEMO result in hyper-IgM syndrome with hypohidrotic ectodermal dysplasia. *Nature Immunol.* **2**, 223–228 (2001).
- Wilkinson, K. D. Regulation of ubiquitin-dependent processes by deubiquitinating enzymes. *FASEB J.* **11**, 1245–1256 (1997).
- Headon, D. J. & Overbeek, P. A. Involvement of a novel TNF receptor homologue in hair follicle induction. *Nature Genet.* **22**, 370–374 (1999).
- Yan, M. et al. Two-amino acid molecular switch in an epithelial morphogen that regulates binding to two distinct receptors. *Science* **290**, 523–527 (2000).
- Hatzivassiliou, E. & Mosialos, G. Cellular signaling pathways engaged by the Epstein-Barr virus transforming protein LMP1. *Front. Biosci.* **7**, d319–d329 (2002).
- Nguyen, L. T. et al. TRAF2 deficiency results in hyperactivity of certain TNFR1 signals and impairment of CD40-mediated responses. *Immunity* **11**, 379–389 (1999).
- Lomaga, M. A. et al. TRAF6 deficiency results in osteopetrosis and defective interleukin-1, CD40, and LPS signaling. *Genes Dev.* **13**, 1015–1024 (1999).
- Yan, M. et al. Identification of a novel death domain-containing adaptor molecule for ectodysplasin-A receptor that is mutated in crinkled mice. *Curr. Biol.* **12**, 409–413 (2002).
- Naito, A. et al. TRAF6-deficient mice display hypohidrotic ectodermal dysplasia. *Proc. Natl Acad. Sci. USA* **99**, 8766–8771 (2002).
- Kaye, K. M. et al. Tumor necrosis factor receptor associated factor 2 is a mediator of NF- κ B activation by latent infection membrane protein 1, the Epstein-Barr virus transforming protein. *Proc. Natl Acad. Sci. USA* **93**, 11085–11090 (1996).
- Schultheiss, U. et al. TRAF6 is a critical mediator of signal transduction by the viral oncogene latent membrane protein 1. *EMBO J.* **20**, 5678–5691 (2001).
- Deng, L. et al. Activation of the I κ B kinase complex by TRAF6 requires a dimeric ubiquitin-conjugating enzyme complex and a unique polyubiquitin chain. *Cell* **103**, 351–361 (2000).
- Wang, C. et al. TAK1 is a ubiquitin-dependent kinase of MKK and IKK. *Nature* **412**, 346–351 (2001).
- Shi, C. S. & Kehrl, J. H. Tumor necrosis factor (TNF)-induced germinal center kinase-related (GCKR) and stress-activated protein kinase (SAPK) activation depends upon the E2/E3 complex Ubc13-Uev1A/TNF receptor-associated factor 2 (TRAF2). *J. Biol. Chem.* **278**, 15429–15434 (2003).
- McKenna, S. et al. An NMR-based model of the ubiquitin-bound human ubiquitin conjugation complex Mms2-Ubc13. The structural basis for lysine 63 chain catalysis. *J. Biol. Chem.* **278**, 13151–13158 (2003).
- Devin, A. et al. The distinct roles of TRAF2 and RIP in IKK activation by TNF-R1: TRAF2 recruits IKK to TNF-R1 while RIP mediates IKK activation. *Immunity* **12**, 419–429 (2000).
- Devin, A. et al. The α and β subunits of I κ B kinase (IKK) mediate TRAF2-dependent IKK recruitment to tumor necrosis factor (TNF) receptor 1 in response to TNF. *Mol. Cell. Biol.* **21**, 3986–3994 (2001).
- Zhang, S. Q., Kovalenko, A., Cantarella, G. & Wallach, D. Recruitment of the IKK signalosome to the p55 TNF receptor: RIP and A20 bind to NEMO (IKK γ) upon receptor stimulation. *Immunity* **12**, 301–311 (2000).
- Hu, M. et al. Crystal structure of a UBP-family deubiquitinating enzyme in isolation and in complex with ubiquitin aldehyde. *Cell* **111**, 1041–1054 (2002).
- Karin, M., Cao, Y., Greten, F. R. & Li, Z. W. NF- κ B in cancer: from innocent bystander to major culprit. *Nature Rev. Cancer* **2**, 301–310 (2002).
- Brummelkamp, T. R., Nijman, S. M. B., Dirac, A. M. G. & Bernards, R. Loss of cylindromatosis tumour suppressor inhibits apoptosis by activating NF- κ B. *Nature* **424**, 797–801 (2003).
- Mosialos, G. et al. The Epstein-Barr virus transforming protein LMP1 engages signaling proteins for the tumor necrosis factor receptor family. *Cell* **80**, 389–399 (1995).
- Mitchell, T. & Sugden, B. Stimulation of NF- κ B-mediated transcription by mutant derivatives of the latent membrane protein of Epstein-Barr virus. *J. Virol.* **69**, 2968–2976 (1995).
- Hatzivassiliou, E., Cardot, P., Zannis, V. I. & Mitsialis, S. A. Ultraspiracle, a *Drosophila* retinoid X receptor α homologue, can mobilize the human thyroid hormone receptor to transactivate a human promoter. *Biochemistry* **36**, 9221–9231 (1997).
- Devergne, O. et al. Association of TRAF1, TRAF2, and TRAF3 with an Epstein-Barr virus LMP1 domain important for B-lymphocyte transformation: role in NF- κ B activation. *Mol. Cell. Biol.* **16**, 7098–7108 (1996).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank A. D'Andrea, Z. J. Chen, C. H. Chung, V. Dixit, M. Hochstrasser, E. Kieff, H. Kikutani, B. Lim, A. Koromilas, H. Nakano, M. Oren and S. Tronick for reagents; and D. Thanos and G. Panayotou for critically reading the manuscript. This work was supported by an International Scholarship from the Howard Hughes Medical Institute, a Human Frontiers Science Program grant and an EMBO Young Investigator award to G.M., and by Cancer Research UK funding to A.A.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.M. (mosialos@fleming.gr).

Loss of the cylindromatosis tumour suppressor inhibits apoptosis by activating NF- κ B

Thijn R. Brummelkamp*, Sebastian M. B. Nijman*, Annette M. G. Dirac* & René Bernards

Division of Molecular Carcinogenesis and Center for Biomedical Genetics, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands

* These authors contributed equally to this work

Protein modification by the conjugation of ubiquitin moieties—ubiquitination—plays a major part in many biological processes, including cell cycle and apoptosis¹. The enzymes that mediate ubiquitin-conjugation have been well-studied, but much less is known about the ubiquitin-specific proteases that mediate de-ubiquitination of cellular substrates^{2,3}. To study this gene family, we designed a collection of RNA interference vectors to suppress 50 human de-ubiquitinating enzymes, and used these vectors to identify de-ubiquitinating enzymes in cancer-relevant pathways. We report here that inhibition of one of these enzymes, the familial cylindromatosis tumour suppressor gene (*CYLD*)⁴, having no known function, enhances activation of the transcription factor NF- κ B. We show that *CYLD* binds to the NEMO (also known as IKK γ) component of the I κ B kinase (IKK) complex, and appears to regulate its activity through de-ubiquitination of TRAF2, as TRAF2 ubiquitination can be modulated by *CYLD*. Inhibition of *CYLD* increases resistance to apoptosis, suggesting a mechanism through which loss of *CYLD* contributes to oncogenesis. We show that this effect can be relieved by aspirin derivatives that inhibit NF- κ B activity⁵, which suggests a therapeutic intervention strategy to restore growth control in patients suffering from familial cylindromatosis.

The family of de-ubiquitinating enzymes (DUBs) consists of ubiquitin carboxy-terminal hydrolases and ubiquitin-specific processing proteases^{1–3}. A role for DUB genes in cancer is suggested by the fact that this family contains both oncogenes^{6,7} and tumour suppressor genes⁴. The strategy we pursued to study the function of the family of DUB enzymes was to inhibit expression of individual family members through RNA interference and search for phenotypes. A total of 50 genes containing the conserved catalytic domain of DUB enzymes could be identified in sequence databases (Fig. 1a, and Supplementary Information). We retrieved the complementary DNA sequences, and selected four unique 19-mer sequences from each transcript for cloning into pSUPER, a vector that mediates suppression of gene expression through synthesis of short hairpin RNAs having small interfering RNA (siRNA)-like properties⁸. We chose to make four knockdown vectors against each DUB to increase the chance that a significant inhibition of DUB expression would be obtained. In total, we made 200 knockdown vectors, which were subsequently pooled into 50 sets of 4 vectors, where each set of vectors was designed to target a single DUB transcript (Fig. 1a).

To ask how effective sets of four knockdown vectors inhibited DUB gene expression, we fused the open reading frame of four of the DUBs to green fluorescent protein (GFP) and determined the levels of GFP–DUB fusion protein expression in the absence and presence of co-expression of the DUB knockdown vectors. A significant reduction in protein levels was induced by all four DUB knockdown vector pools, whereas control p21–red fluorescent protein (RFP) fusion protein was unaffected (Fig. 1b). We conclude that this strategy allows efficient inhibition of DUB expression.

To study members of the DUB family, we asked if suppression of

DUBs affected the activity of NF- κ B, a transcription factor involved in inflammation, immune responses and protection against apoptosis⁹. We transfected an NF- κ B-luciferase reporter gene, together with each of the 50 sets of 4 DUB knockdown vectors into human U2-OS cells, and measured the effect of DUB knockdown on tumour necrosis factor- α (TNF- α)-activation of NF- κ B. Only one of the sets of DUB knockdown vectors (no. 36) enhanced TNF- α -activation of NF- κ B (Fig. 1c). Importantly, the DUB no. 36 set of knockdown vectors targets the cylindromatosis tumour suppressor gene *CYLD*⁴ (see Supplementary Information). Levels of haemagglutinin (HA)–*CYLD* protein levels were significantly reduced by the most active of the four knockdown vectors (pSUPER–*CYLD*) (Fig. 1d). *CYLD* knockdown did not enhance basal level of NF- κ B activity, but caused a further increase in NF- κ B activation by phorbol 12-myristate 13-acetate (PMA) (Fig. 1e). This indicates that *CYLD* loss activates NF- κ B in response to diverse stimuli. Together, these data suggest that *CYLD* is an NF- κ B regulator.

NF- κ B is inactivated by I κ B proteins. Signalling through the I κ B kinase (IKK) complex, containing the I κ B kinases IKK α and β and the structural component NEMO (or IKK γ), causes phosphorylation and subsequent degradation of I κ B, allowing activation of NF- κ B^{9,10}. The observed effect of *CYLD* on NF- κ B activity could result from an effect of *CYLD* on the TNF- α receptor, the TNF-receptor-associated factor (TRAF) proteins, a downstream effect on the IKK complex or on the I κ B/NF- κ B complex. To address this, we asked if *CYLD* associates with members of the NF- κ B signalling machinery. We found that *CYLD* co-immunoprecipitated with NEMO (Fig. 2a). Since *CYLD* interacts with NEMO in a yeast two hybrid assay (see accompanying manuscript¹¹), this is likely to be direct. However, owing to the lower expression of I κ B α and IKK β (Fig. 2a), we cannot exclude the possibility that *CYLD* also interacts with other components of the IKK complex.

To study the effect of *CYLD* on IKK activity, we measured IKK β activity in an *in vitro* kinase assay. In the absence of TNF- α no IKK β kinase activity towards I κ B α could be detected (Fig. 2b). As expected, TNF- α treatment stimulated IKK β kinase activity. Importantly, this activity was enhanced when cells were co-transfected with pSUPER–*CYLD* (compare Fig. 2b, lanes 5, 6). No effects were seen of *CYLD* knockdown on IKK β protein levels (Fig. 2b, lower panel), suggesting that *CYLD* does not regulate IKK β abundance. Consistent with increased IKK β kinase activity by pSUPER–*CYLD*, we observed that *CYLD* knockdown resulted in a reduction in I κ B α levels, an endogenous substrate of IKK β kinase (Fig. 2c). This decrease in I κ B α level occurs without TNF- α treatment, indicating that there might be some IKK activity in U2-OS cells under the tissue culture conditions used, or that *CYLD* is required to maintain shut-off of IKK signalling in the absence of activating signals. These data indicate that *CYLD* acts as an antagonist of the IKK complex that binds to the NEMO component, and that a reduction in *CYLD* stimulates IKK-activity. A prediction from these data is that expression of a catalytically inactive mutant of *CYLD* will also cause a reduction of the I κ B α levels through competition with endogenous *CYLD*. Figure 2d shows that transfection of cells with a point mutant of *CYLD* in which the active site cysteine is mutated⁷ causes a similar reduction in I κ B α levels as seen with pSUPER–*CYLD* (compare Fig. 2d, lanes 2 and 4). We conclude that the de-ubiquitinating enzyme activity of *CYLD* is required to act on NF- κ B.

TNF receptor-associated factors (TRAFs) are involved in cytokine-mediated activation of NF- κ B^{12,13}. NF- κ B-inducing activity of TRAF2 and TRAF6 requires their ubiquitin ligase activity and non-destructive lysine-63 poly-ubiquitination^{14,15}. Their ubiquitination during TNF signalling makes these proteins candidate substrates for *CYLD*. We co-transfected TRAF2 and HA-ubiquitin expression vectors, and monitored levels of TRAF2 ubiquitination in the presence of both wild-type *CYLD* and an active site mutant

of CYLD. Figure 2e shows that TRAF2 is ubiquitinated, in agreement with the notion that TRAF2 over-expression leads to activation of NF- κ B¹¹. Importantly, co-expression of wild-type CYLD caused a reduction in TRAF2 ubiquitination, and an active site mutant of CYLD markedly increased levels of ubiquitinated TRAF2 (Fig. 2e). Although these results do not rule out the possibility that CYLD could have other targets in the pathway, or that the effects seen on TRAF2 ubiquitination are indirect, they are consistent with a model in which the effect of CYLD on NF- κ B activation is mediated by CYLD-dependent de-ubiquitination of TRAF2.

When combined with inhibitors of transcription or translation, TNF- α can be a potent inducer of apoptosis^{16–19}. This pro-apoptotic

activity of TNF- α can be inhibited by activation of NF- κ B, which activates a number of anti-apoptotic genes²⁰. Since CYLD knockdown stimulates PMA-induced activation of NF- κ B, we asked if CYLD and PMA also collaborate to inhibit TNF- α -induced apoptosis. HeLa cells were treated with TNF- α in the presence of cycloheximide (CHX) to induce apoptosis both with and without pre-treatment with PMA, which stimulates expression of anti-apoptotic NF- κ B target genes (see Methods). Figure 3a, b shows that treatment with TNF- α efficiently induced apoptosis in some 95% of the HeLa cells, as judged by TUNEL assay (data not shown). As expected, pre-treatment with PMA resulted in an approximately four-fold increase of the number of viable cells. Significantly, PMA pre-treatment in CYLD knockdown cells

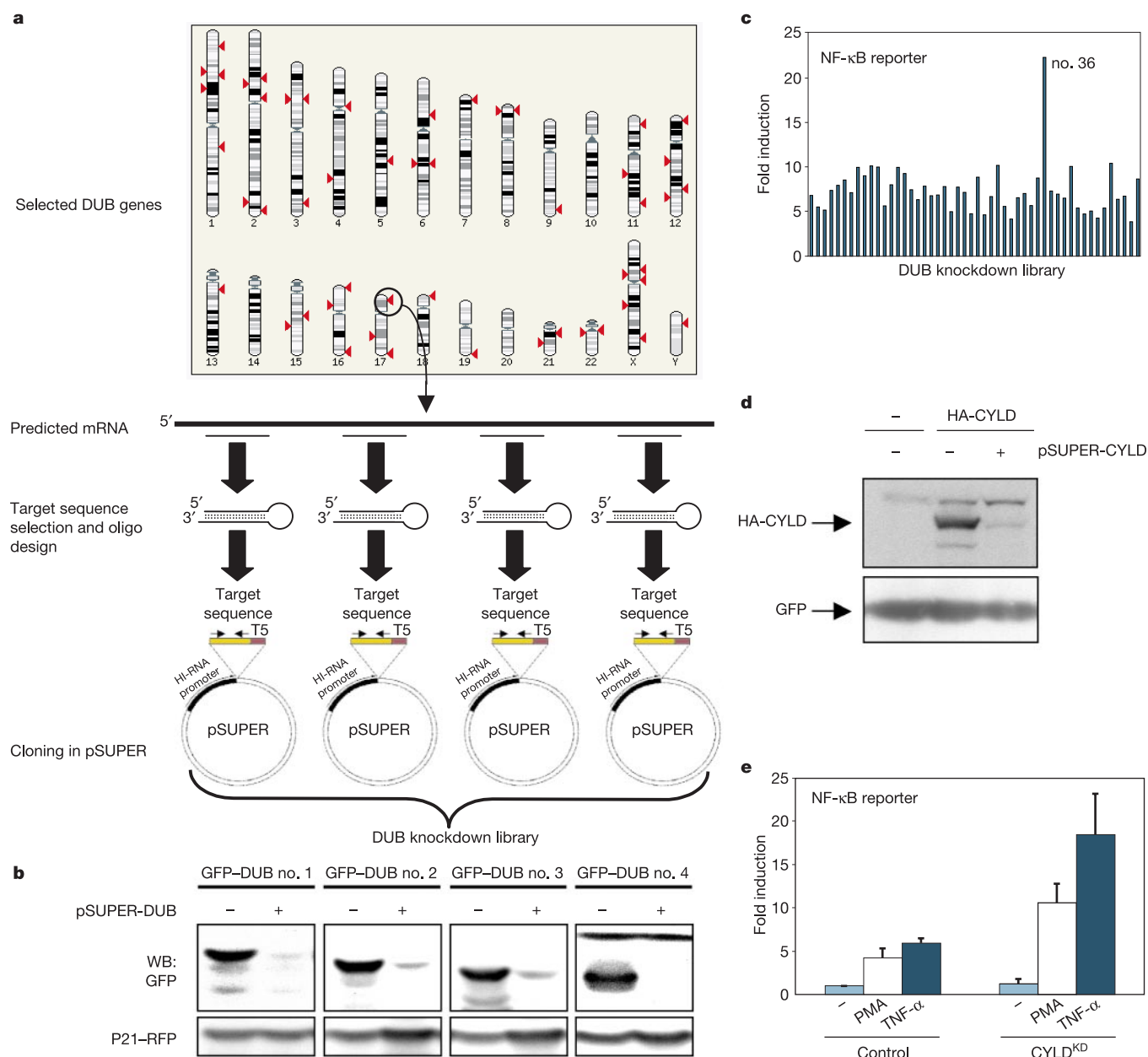


Figure 1 Genome-wide DUB knockdown screen. **a**, Overview of identified DUBs using the Ensembl database, and construction of the knockdown library. **b**, HEK293 cells were transfected and immunoblotted against GFP or p21-RFP (control). WB, western blot. **c**, U2-OS cells were co-transfected with a NF- κ B luciferase reporter and individual members of the DUB knockdown library. Cells were stimulated with TNF- α and luciferase

activity was measured. **d**, U2-OS cells were transfected with HA-tagged CYLD and pSUPER-CYLD or empty vector. Extracts were immunoblotted against HA or GFP (control). **e**, U2-OS cells were transfected as in **c**, and were stimulated with PMA or TNF- α . Average values are of three independent transfections ($\pm$ s.d.). CYLD^{KD}, CYLD knockdown.

resulted in an even larger fraction of surviving cells, suggesting that CYLD loss can be anti-apoptotic (Fig. 3a, b), probably through enhanced activation of NF- κ B. Consistent with this, CYLD knockdown and PMA also collaborated in NF- κ B activation in HeLa cells (Fig. 3c).

NF- κ B can be inhibited by a number of pharmacological agents, including sodium salicylate (an aspirin derivative) and prostaglandin A1 (PGA1)^{5,21}. Both inhibit IKK β , the kinase which is hyper-activated as a result of CYLD knockdown (Fig. 2b). To address whether the effect of CYLD knockdown on NF- κ B activation can be counteracted by sodium salicylate or PGA1, we transfected U2-OS cells with the NF- κ B-reporter and stimulated NF- κ B with PMA. As observed before, pSUPER-CYLD enhanced PMA-stimulated NF- κ B activity (Fig. 4a, b). This effect of CYLD knockdown on NF- κ B could be suppressed by both sodium salicylate and PGA1 (Fig. 4a, b), indicating that these compounds can compensate for CYLD suppression in this assay.

It is likely that loss of CYLD confers resistance to apoptosis through activation of NF- κ B. If correct, the protective effect of

CYLD knockdown could be reversed by the NF- κ B inhibitor sodium salicylate. We tested this by treating HeLa cells with PMA in the presence or absence of pSUPER-CYLD. Apoptosis was induced by the addition of TNF- α and CHX. CYLD knockdown and PMA treatment again conferred resistance to TNF- α -induced apoptosis (Fig. 4c). Exposure of cells to sodium salicylate abolished the protective effect of CYLD knockdown on apoptosis (Fig. 4c). This

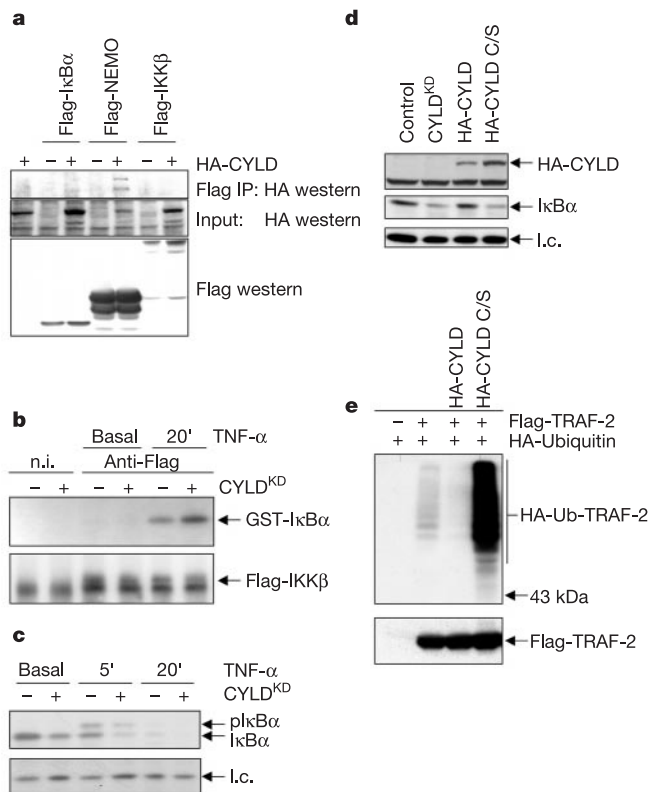


Figure 2 CYLD is an antagonist of NF- κ B signalling. **a**, HEK293 cells were transfected as indicated, and immunoprecipitated (IP) using Flag antibody. IPs were immunoblotted for HA-tagged CYLD (upper panel), whole cell extracts were immunoblotted for Flag-tagged IκBα, IKKβ and NEMO (lower panel) and for HA-tagged CYLD (middle panel). **b**, Flag-IKKβ was immunoprecipitated with either non-immune serum (n.i.) or Flag antibody from both untreated cells (basal) or cells treated for 20 min. with TNF- α (20') and lysates were incubated with GST-IκBα (1-72) in the presence of ³²P-γATP (upper panel). Immunoprecipitated IKKβ was visualized by immunoblotting with Flag antibody (lower panel). **c**, U2-OS cells electroporated with pSUPER-CYLD or empty vector were stimulated with TNF- α and immunoblotted for IκBα (I.c., loading control). **d**, U2-OS cells were transfected and immunoblotted for IκBα. **e**, HEK293 cells were transfected and TRAF2 was immunoprecipitated using Flag antibodies. Immunoprecipitates were immunoblotted for HA-tagged ubiquitin (upper panel) and whole cell extracts for Flag-TRAF2 (lower panel).

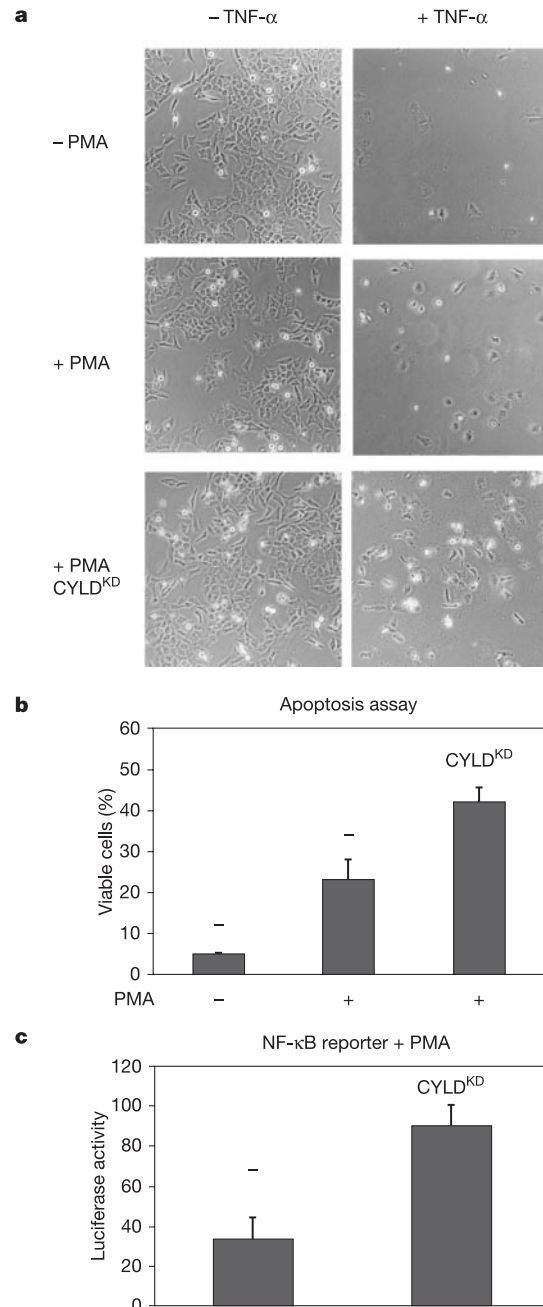


Figure 3 CYLD loss protects against TNF- α induced apoptosis. **a**, **b**, HeLa cells were electroporated as indicated. To stimulate NF- κ B, cells were pre-treated with PMA for 2 hours or left untreated prior to inducing apoptosis by the simultaneous addition of TNF- α and CHX. Photos were taken 16 hours after TNF- α /CHX addition (**a**) and percentage of viable cells was quantified. The average of three independent experiments ($\pm$ s.d.) is shown (**b**). **c**, HeLa cells were transfected with a NF- κ B luciferase reporter and pSUPER-CYLD or empty vector. Cells were stimulated with PMA for two hours and luciferase activity was measured. The average value of three independent transfections ($\pm$ s.d.) is shown.

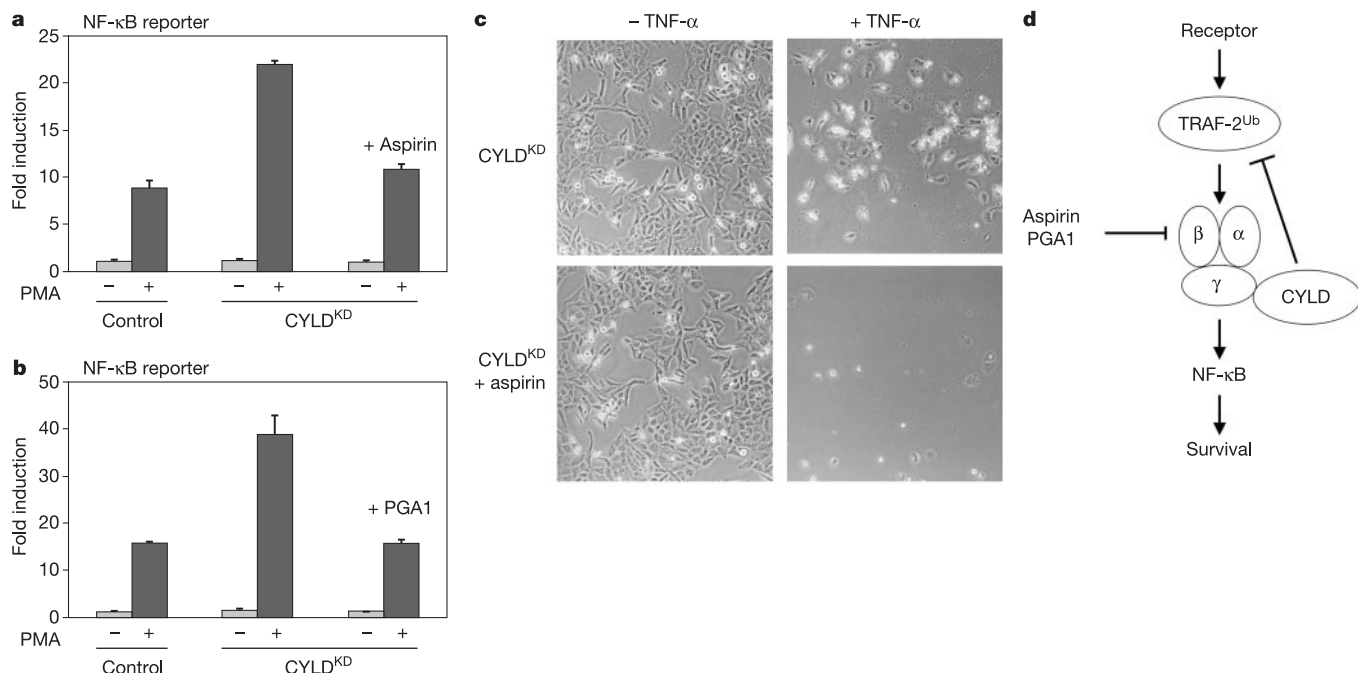


Figure 4 Anti-apoptotic effects of CYLD loss can be reversed by aspirin. **a, b**, U2-OS cells were transfected as in Fig. 1c, and either left untreated, stimulated with PMA, PMA and sodium salicylate (**a**, labelled 'aspirin') or PMA and prostaglandin A1 (**b**) and luciferase activity was measured. The average values of three independent transfections

($\pm$ s.d.) are shown. **c**, Electroporated HeLa cells were pre-treated with PMA or PMA and sodium salicylate (labelled 'aspirin') for two hours. Apoptosis was induced by addition of TNF- α and CHX. Photos were taken 16 hours later. **d**, Model of the consequences of CYLD loss on NF- κ B activity and cell survival.

indicates that CYLD loss is anti-apoptotic through the activation of IKK β and subsequently NF- κ B.

We describe here a high-throughput RNA interference screen in mammalian cells to identify novel regulators of NF- κ B. We focused on ubiquitin-specific processing proteases, as these proteins are antagonists of the ubiquitin conjugating enzymes and ubiquitin ligases. We identified the familial cylindromatosis tumour suppressor gene (*CYLD*) as a novel negative regulator of NF- κ B, thus establishing the first (to our knowledge) direct link between the NF- κ B signalling cascade and a tumour suppressor gene. We propose a model in which CYLD functions to turn TNF-receptor signalling off by removing ubiquitin from TRAF2 (Fig. 4d). Although CYLD may have other targets than NF- κ B, we suggest that the deregulated proliferation in the skin appendages in familial cylindromatosis results from a mild shift in the balance between proliferation and apoptosis in favour of proliferation. This is supported by the observation that most cylindromas have a diploid karyotype and are rarely metastatic⁴.

In the vast majority of cell types, activation of NF- κ B is an oncogenic event⁹. However, it has recently been shown that inhibition of NF- κ B in keratinocytes contributes to neoplasia²². Importantly, cylindromas are not derived from keratinocytes but from hair follicles and exocrine glands²³. These skin appendages require NF- κ B activity, as mice with suppressed NF- κ B have defects in the development of these tissues due to increased apoptosis^{24,25}.

Our observation that enhanced protection from apoptosis resulting from *CYLD* suppression can be reversed by simple pharmacological agents like sodium salicylate and prostaglandin A1 suggests a strategy to restore normal growth control in patients suffering from familial cylindromatosis. We are currently testing this in a clinical trial using topical application of aspirin derivatives on cylindromas.

Methods

Materials, antibodies, and plasmid construction

To generate DUB knockdown vectors, four annealed sets of oligonucleotides encoding short hairpin transcripts corresponding to one DUB enzyme (see Supplementary Information) were cloned individually into pSUPER. Bacterial colonies were pooled and used for plasmid preparation. To generate GFP-DUB fusion proteins, the corresponding DUB enzymes were PCR amplified and cloned into pEGFP-N1 (Clontech). pNF- κ B-Luc vector was obtained from Clontech, SV40-Renilla from Promega. PMA, TNF- α , prostaglandin A1 and cycloheximide were purchased from Sigma. HA-tagged CYLD was PCR amplified, and cloned into pcDNA 3.1 (-), Flag tagged NEMO was generated by cloning an *EcoRI-XbaI* NEMO containing fragment into pcDNA-flag. Anti-I κ B α (c-21) and HA tag (Y-11) antibodies were obtained from Santa Cruz, anti-Flag M2 from Sigma and anti-GFP rabbit polyclonal serum was provided by J. Neefjes.

Cell cultures, transient transfections and reporter assays

All cells were cultured in Dulbecco's modified Eagle medium (DMEM) supplemented with 10% fetal calf serum. High efficiency electroporation of cells was done as described²⁶. Reporter assays were carried out using calcium phosphate transfection of 0.5 μ g NF- κ B-Luc, 1 ng SV40-Renilla and 2.5 μ g pSUPER vectors. Forty-eight hours after transfection, cells were stimulated with 200 nM PMA or 20 ng ml⁻¹ TNF- α and luciferase activity was measured 72 hours after transfection. Sodium salicylate (10 mM) or prostaglandin A1 (8 μ M) was added to the cells 48 hours after transfection, and reporter activity was measured 72 hours after transfection. In HeLa cells, NF- κ B activity was measured 2 hours after PMA stimulation.

Immunoblotting, immunoprecipitation and kinase assay

Western blots were performed using whole cell extracts, separated on 8–12% SDS-PAGE gels and transferred to polyvinylidene difluoride membranes (Millipore). Western blots were probed with the indicated antibodies. HEK293 cells were transfected by calcium phosphate precipitation with the indicated plasmids; 48 hours after transfection cells were lysed in ELB buffer (0.25 M NaCl, 0.1% NP-40, 50 mM HEPES pH 7.3) supplemented with 'Complete' protease inhibitors (Roche), and protein complexes were immunoprecipitated with 2 μ g of the indicated antibodies conjugated to protein G sepharose beads. To detect TRAF2 ubiquitination, immunoprecipitations were performed using RIPA buffer (50 mM Tris pH 8.0, 150 mM NaCl, 1% NP-40, 0.5% deoxycholic acid and 0.1% SDS). Immunoprecipitation/kinase assays were performed essentially as described²⁷.

Apoptosis assays

Experimental design. In the experiments described in Figs 3 and 4, we first treat cells with

PMA or PMA in the presence of CYLD knockdown for 2–3 hours. This allows NF- κ B to become active and stimulate expression of the anti-apoptotic NF- κ B target genes. Then we add TNF- α and CHX at the same time. The CHX is added because TNF- α has two effects; it induces apoptosis in a transcription-independent fashion, and induces NF- κ B to protect from apoptosis, which requires transcription. By adding CHX and TNF- α at the same time, the anti-apoptotic effects of TNF- α are blocked (by CHX), leaving only the apoptotic signalling pathway intact. This is why combined treatment with TNF- α and CHX is such a powerful inducer of apoptosis. Thus, in this experimental set-up the apoptotic effects of TNF- α can only be rescued by activation of NF- κ B target genes prior to CHX treatment (that is, through the combined effects of PMA + CYLD knockdown).

Seventy-two hours after electroporation of HeLa cells with the indicated plasmids, cells were treated with 200 nM PMA for 2–3 hours to stimulate NF- κ B activity, followed by a 12-hour incubation in medium containing both 10 ng ml⁻¹ TNF- α and 10 μ g ml⁻¹ CHX to induce apoptosis. Viable cells were quantified using the Trypan blue exclusion method. To inhibit NF- κ B activity, medium was supplemented with 10 mM sodium salicylate 3.5 hours before TNF- α addition.

Received 12 March; accepted 8 May 2003; doi:10.1038/nature01811.

1. Wilkinson, K. D. Ubiquitination and deubiquitination: Targeting of proteins for degradation by the proteasome. *Semin. Cell Dev. Biol.* **11**, 141–148 (2000).
2. Chung, C. H. & Baek, S. H. Deubiquitinating enzymes: Their diversity and emerging roles. *Biochem. Biophys. Res. Commun.* **266**, 633–640 (1999).
3. D'Andrea, A. & Pellman, D. Deubiquitinating enzymes: A new class of biological regulators. *Crit. Rev. Biochem. Mol. Biol.* **33**, 337–352 (1998).
4. Bignell, G. R. *et al.* Identification of the familial cylindromatosis tumour-suppressor gene. *Nature Genet.* **25**, 160–165 (2000).
5. Yin, M. J., Yamamoto, Y. & Gaynor, R. B. The anti-inflammatory agents aspirin and salicylate inhibit the activity of I κ B kinase- β . *Nature* **396**, 77–80 (1998).
6. Nakamura, T., Hillova, J., Mariage-Samson, R. & Hill, M. Molecular cloning of a novel oncogene generated by DNA recombination during transfection. *Oncogene Res.* **2**, 357–370 (1988).
7. Papa, F. R. & Hochstrasser, M. The yeast *DOA4* gene encodes a deubiquitinating enzyme related to a product of the human *tre-2* oncogene. *Nature* **366**, 313–319 (1993).
8. Brummelkamp, T. R., Bernards, R. & Agami, R. A system for stable expression of short interfering RNAs in mammalian cells. *Science* **296**, 550–553 (2002).
9. Karin, M., Cao, Y., Greten, F. R. & Li, Z. W. NF- κ B in cancer: From innocent bystander to major culprit. *Nature Rev. Cancer* **2**, 301–310 (2002).
10. Smahi, A. *et al.* The NF- κ B signalling pathway in human diseases: From incontinentia pigmenti to ectodermal dysplasias and immune-deficiency syndromes. *Hum. Mol. Genet.* **11**, 2371–2375 (2002).
11. Trompouki, E. *et al.* CYLD is a deubiquitinating enzyme that negatively regulates NF- κ B activation by TNFR family members. *Nature* **424**, 793–796 (2003).
12. Bradley, J. R. & Pober, J. S. Tumor necrosis factor receptor-associated factors (TRAFs). *Oncogene* **20**, 6482–6491 (2001).
13. Chung, J. Y., Park, Y. C., Ye, H. & Wu, H. All TRAFs are not created equal: Common and distinct molecular mechanisms of TRAF-mediated signal transduction. *J. Cell Sci.* **115**, 679–688 (2002).
14. Shi, C. S. & Kehrl, J. H. TNF-induced GCKR and SAPK activation depends upon the E2/E3 complex Ubc13-Uev1A/TRAFF2. *J. Biol. Chem.* **278**, 15429–15434 (2003).
15. Wang, C. *et al.* TAK1 is a ubiquitin-dependent kinase of MKK and IKK. *Nature* **412**, 346–351 (2001).
16. Holtmann, H., Hahn, T. & Wallach, D. Interrelated effects of tumor necrosis factor and interleukin 1 on cell viability. *Immunobiology* **177**, 7–22 (1988).
17. Wang, C. Y., Mayo, M. W. & Baldwin, A. S. Jr TNF- and cancer therapy-induced apoptosis: Potentiation by inhibition of NF- κ B. *Science* **274**, 784–787 (1996).
18. Liu, Z. G., Hsu, H., Goeddel, D. V. & Karin, M. Dissection of TNF receptor 1 effector functions: JNK activation is not linked to apoptosis while NF- κ B activation prevents cell death. *Cell* **87**, 565–576 (1996).
19. De Smaele, E. *et al.* Induction of gadd45 β by NF- κ B downregulates pro-apoptotic JNK signalling. *Nature* **414**, 308–313 (2001).
20. Baud, V. & Karin, M. Signal transduction by tumor necrosis factor and its relatives. *Trends Cell Biol.* **11**, 372–377 (2001).
21. Rossi, A. *et al.* Anti-inflammatory cyclopentenone prostaglandins are direct inhibitors of I κ B kinase. *Nature* **403**, 103–108 (2000).
22. Dajee, M. *et al.* NF- κ B blockade and oncogenic Ras trigger invasive human epidermal neoplasia. *Nature* **421**, 639–643 (2003).
23. van Balkom, I. D. & Hennekam, R. C. Dermal eccrine cylindromatosis. *J. Med. Genet.* **31**, 321–324 (1994).
24. Schmidt-Ullrich, R. *et al.* NF- κ B activity in transgenic mice: Developmental regulation and tissue specificity. *Development* **122**, 2117–2128 (1996).
25. Schmidt-Ullrich, R. *et al.* Requirement of NF- κ B/Rel for the development of hair follicles and other epidermal appendages. *Development* **128**, 3843–3853 (2001).
26. Agami, R. & Bernards, R. Distinct initiation and maintenance mechanisms cooperate to induce G1 cell cycle arrest in response to DNA damage. *Cell* **102**, 55–66 (2000).
27. Chen, G., Cao, P. & Goeddel, D. V. TNF-induced recruitment and activation of the IKK complex require Cdc37 and Hsp90. *Mol. Cell* **9**, 401–410 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank S. Lens, A. Lund and J. Borst for reagents, and M. Madiredjo for assistance. This work was supported by the Centre for Biomedical Genetics (CBG) and the Netherlands Organization for Scientific Research (NWO). A.D. was supported by a long-term fellowship from EMBO.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.B. (r.bernards@nki.nl).

The tumour suppressor CYLD negatively regulates NF- κ B signalling by deubiquitination

Andrew Kovalenko^{1,2}, Christine Chable-Bessia²,
Giuseppina Cantarella^{1,3}, Alain Israël², David Wallach¹ & Gilles Courtis²

¹Department of Biological Chemistry, The Weizmann Institute of Science, 76100 Rehovot, Israel

²Unité de Biologie Moléculaire de l'Expression Génique, CNRS URA 2582, Institut Pasteur, 75015 Paris, France

³Department of Experimental and Clinical Pharmacology, University of Catania School of Medicine, I-95125 Catania, Italy

NF- κ B transcription factors have key roles in inflammation, immune response, oncogenesis and protection against apoptosis^{1,2}. In most cells, these factors are kept inactive in the cytoplasm through association with I κ B inhibitors. After stimulation by various reagents, I κ B is phosphorylated by the I κ B kinase (IKK) complex³ and degraded by the proteasome, allowing NF- κ B to translocate to the nucleus and activate its target genes. Here we report that CYLD, a tumour suppressor that is mutated in familial cylindromatosis⁴, interacts with NEMO, the regulatory subunit of IKK^{5,6}. CYLD also interacts directly with tumour-necrosis factor receptor (TNFR)-associated factor 2 (TRAF2), an adaptor molecule involved in signalling by members of the family of TNF/nerve growth factor receptors. CYLD has deubiquitinating activity that is directed towards non-K48-linked polyubiquitin chains, and negatively modulates TRAF-mediated activation of IKK, strengthening the notion that ubiquitination is involved in IKK activation by TRAFs and suggesting that CYLD functions in this process. Truncations of CYLD found in cylindromatosis result in reduced enzymatic activity, indicating a link between impaired deubiquitination of CYLD substrates and human pathophysiology.

The regulatory subunit of IKK, NEMO (also known as IKK- γ), has an essential role in NF- κ B activation and its dysfunction is associated with several distinct human pathologies⁷. On applying the carboxy-terminal half of NEMO as a bait in two-hybrid screening, we found that it bound specifically to the tumour suppressor CYLD (Fig. 1a). CYLD was originally identified through positional cloning of the gene responsible for familial cylindromatosis, an autosomal dominant disease characterized by benign tumours derived from cells of skin appendages, which appear predominantly in hairy areas of the body during adulthood⁴.

The binding site for CYLD on NEMO was found to correspond to the last 39 amino acids of the NEMO protein, where a zinc-finger is located (Fig. 1a). In further two-hybrid tests, CYLD was found to bind to TRAF2 but not to other members of the TRAF family or to any of the following signalling proteins: TRADD, RIP, RIP2, NIK, FADD, cFLIP, caspase 8, APAF1, IAP-1, the p55 TNFR intracellular domain, FAS-IC or A20 (data not shown). The whole TRAF domain of TRAF2 seemed to be required for the CYLD–TRAF2 interaction (Fig. 1b). We found that the NEMO-binding site is located between amino acids 470 and 684 in CYLD, whereas the TRAF2-binding site is located between CYLD amino acids 394 and 470. Because TRAF2 binds to the sequence motif Pro-X-Gln-X-(Ser/Thr)⁸ and we found a similar sequence in the TRAF2-binding domain of CYLD (Pro 453-Val 454-Gln 455-Glu 456-Ser 457), we examined the participation of this sequence in the TRAF2–CYLD interaction by mutating Ser 457 to alanine. This mutation abolished TRAF2 binding to CYLD (Fig. 1c).

CYLD binding to NEMO and to TRAF2 could be also detected in transfected human embryonic kidney (HEK) 293T cells (Fig. 2a, d).

Both TRAF2 and TRAF6 have been shown to act as E3 ubiquitin ligases and to ubiquitinate themselves through K63-Ub conjugation^{11–13}. This modification, in contrast to K48-Ub conjugation, is not associated with proteolysis, but instead triggers signalling through an undefined mechanism¹². When we co-transfected TRAF2 or TRAF6 with HA-ubiquitin into HEK 293T cells, polyubiquitination of both proteins was indeed observed (Fig. 3c).

Transfection of CYLD together with NEMO, TRAF2 or TRAF6 resulted in disappearance of the ubiquitinated forms of the three proteins (Fig. 3c). Notably, co-transfection of TRAF2 with a Ser 457-to-alanine mutant of CYLD, (S/A)CYLD, that could not bind TRAF2 resulted in a lower degree of deubiquitination as compared with cells co-transfected with TRAF2 and wild-type CYLD, indicating that the specific binding of TRAF2 to CYLD assists its deubiquitination (Fig. 3d). By contrast, (S/A)CYLD was just as potent as the wild-type protein in its ability to act on either TRAF6 or NEMO (Fig. 3d). The truncated version of CYLD originally isolated during our two-hybrid screen (clone 10 or (Δ N393)CYLD; Fig. 1c) showed a restricted substrate specificity. It could deubiquitinate TRAF2 and NEMO, but not TRAF6 (Supplementary Fig. 3). CYLD also showed deubiquitinating activity in a cell-free assay (Fig. 3e). Replacement of the putative catalytic histidine residue of CYLD (Fig. 3a) by an asparagine residue, (H/N)CYLD, resulted in total loss of its activity, showing that the enzymatic activity of CYLD, rather than interference caused by binding to NEMO or TRAF2, was responsible for the observed effect (Fig. 3c, e). CYLD did not affect the K48-linked ubiquitination of β -catenin or I κ B α (refs 1, 14, and Fig. 3f), further indicating that it functions as a K63-Ub-specific deubiquitinase.

As discussed above, the shortest truncation of CYLD that generates cylindromatosis (introduction of a stop codon at Arg 936) eliminates subdomain III of the histidine box⁴. We therefore examined the impact of this mutation, (R/X)CYLD, on the enzymatic activity of CYLD. We observed a severe reduction in deubiquitinase activity of (R/X)CYLD both in transfected cells and in a cell-free assay (Fig. 3c, e). This confirms the role of this domain

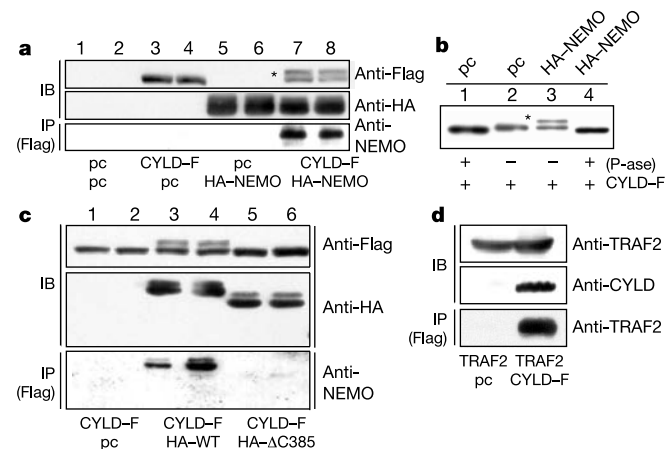


Figure 2 *In vivo* interaction of CYLD with NEMO and TRAF2. **a**, Interaction between CYLD and NEMO. HEK 293T cells were transfected with the indicated combinations of empty vector (pc) and expression constructs encoding HA–NEMO or CYLD–Flag (CYLD-F). Cell extracts were analysed by immunoblotting either directly (IB) or after immunoprecipitation with antibodies against Flag (IP). Asterisk indicates a retarded CYLD species. **b**, NEMO–CYLD interaction results in phosphorylation of CYLD. Lysates of HEK 293T cells transfected with the indicated constructs were incubated without or with λ -phosphatase (P-ase), and then analysed directly by immunoblotting with antibodies against Flag. **c**, The zinc-finger of NEMO is required for interaction with CYLD. The experiment was done as in **a**; HA–WT, full-length NEMO; HA– Δ C385, NEMO lacking the zinc-finger. **d**, Interaction between CYLD and TRAF2. Extracts were analysed by immunoblotting either directly (IB) or after immunoprecipitation with antibodies against Flag (IP).

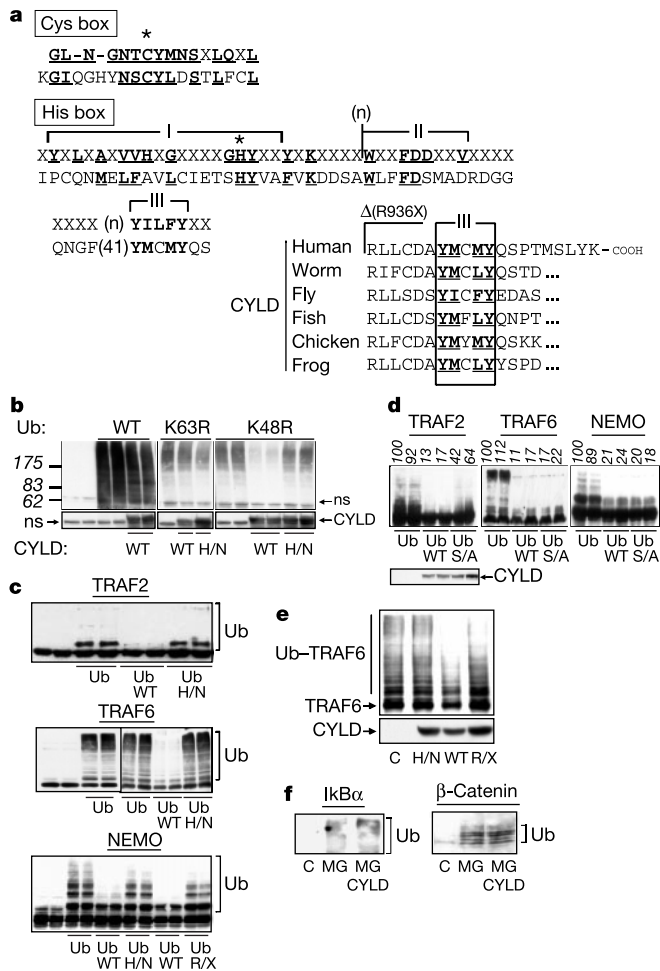


Figure 3 CYLD is an active UBP deubiquitinase with restricted substrate specificity. **a**, Top, alignment of CYLD fragments (lower row) with the consensus of the UBP family, and conserved residues in UBP proteins (see the Protein families database of alignments and HMMs (Pfam) at <http://www.sanger.ac.uk/Software/Pfam>). Asterisks indicate cysteine and histidine residues implicated in the catalytic mechanism, n indicates a variation in amino-acid number between UBP proteins. Bottom, degree of evolutionary conservation in the C-terminal region of CYLD corresponding to block III of the histidine box, and the site of the most distal mutation (R936X) found in cylindromatosis (family TT17), which results in truncation of the last 20 residues of CYLD⁴. **b**, CYLD is a deubiquitinating enzyme affecting non-K48-linked ubiquitination. HA-tagged human wild-type, K48R or K63R ubiquitin was co-transfected with empty vector or a construct expressing wild-type CYLD (WT) or catalytically inactive CYLD (H/N). Overall protein ubiquitination (top) and the expression of CYLD (bottom) were assessed by immunoblotting with antibodies against HA and Flag, respectively. ns, nonspecific signal. Results are representative of three independent experiments exposed to achieve similar signal intensities. **c**, CYLD can deubiquitinate TRAF6, TRAF2 and NEMO *in vivo*. HEK 293T cells were transfected with pCMV-Flag-TRAF6, pCMV-HA-NEMO or pCMV-HA-TRAF2, plus empty vector or a pCMV-HA-ubiquitin (Ub) plasmid. The effect of wild-type CYLD, His-box-mutated CYLD (H/N), or C-terminally truncated CYLD (R/X) on ubiquitinated molecules was assessed by immunoblotting with antibodies against TRAF6, NEMO or TRAF2. **d**, TRAF2 deubiquitination by CYLD is controlled by their specific interaction. Conditions were similar to those in **c**. S/A is the Flag-tagged S457A CYLD mutant. Numbers indicate the relative extent of ubiquitination. The relative expression of transfected CYLD constructs is shown below. **e**, CYLD has deubiquitination activity in a cell-free assay. Equal amounts of wild-type CYLD or the two CYLD mutants (H/N, R/X) obtained from transfected cell lysates were incubated with Flag-TRAF6 beads in deubiquitination buffer (Methods). TRAF6 ubiquitination was analysed with antibodies against TRAF6. **f**, CYLD does not affect the ubiquitination of IκBα or β-catenin. The analysis was done essentially as in **c**, except that cells were pretreated with the proteasomal inhibitor MG132 (MG) to allow ubiquitinated molecules to accumulate for detection. IκBα ubiquitination was induced by co-transfecting cells with IKK2(E/E)¹⁷ and detected with antibodies against IκBα. Ubiquitination of β-catenin was detected with antibodies against β-catenin after immunoprecipitation with antibodies against HA.

in the deubiquitinase activity and suggests that cylindromatosis pathology results from defective deubiquitination of *in vivo* CYLD substrates.

We assessed the affect of overexpressing CYLD on TNF and interleukin-1 (IL-1) function and found that it inhibited the induction of NF-κB by both cytokines (Fig. 4a). This inhibition was diminished when the enzymatically inactive mutant (H/N)CYLD was used instead of wild-type CYLD (Fig. 4a). Overexpression of the mutant (S/A)CYLD, which does not bind TRAF2, also decreased the inhibitory effect of CYLD on NF-κB activation induced by TNF, which involves TRAF2, but not its inhibitory effect on NF-κB activation induced by IL-1, which involves TRAF6 (Fig. 4a).

Overexpression of CYLD also substantially inhibited NF-κB activation induced by several members of the TNFR and IL-1 receptor/Toll-like receptor families, including TNFR-1, EDAR and TLR4 (Fig. 4b), as well as by LMP1, TRAF2, TRAF5, TRAF6, RIP, TRADD and Myd88, of which the effect on TRAF2 was the strongest (Fig. 4c). (H/N)CYLD was only slightly inhibitory to these signalling proteins, whereas (ΔN393)CYLD strongly inhibited TRAF2- but not TRAF6-dependent activation of NF-κB (Supplementary Fig. 3).

Notably, when NF-κB was activated by NIK (a kinase that binds to and activates IKK1; ref. 15), by Tax (a viral protein that directly interacts with NEMO and activates IKK¹⁶), or by a constitutively active form of IKK2 (IKK2(E/E)¹⁷), the overexpressed CYLD was barely inhibitory and no difference was seen between wild-type and

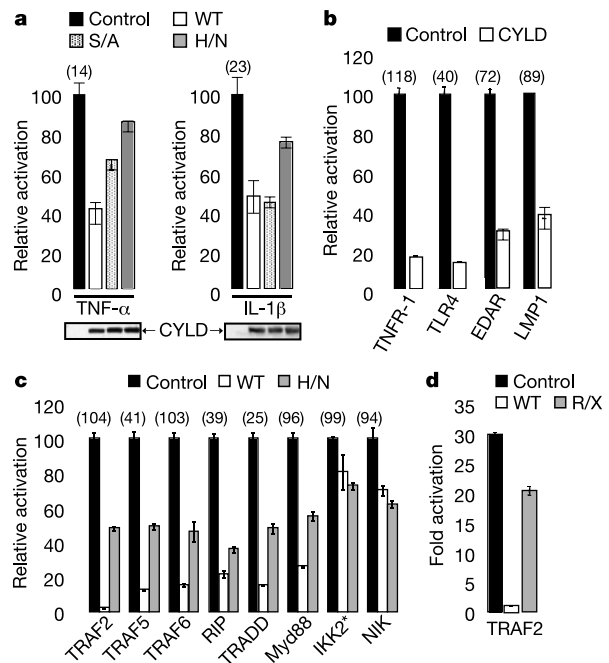


Figure 4 Overexpression of CYLD inhibits NF-κB activation and this inhibition requires its catalytic activity. **a**, CYLD-mediated inhibition of NF-κB activation by TNF-α or IL-1β. HEK 293T cells were co-transfected with either empty vector (control) or constructs expressing wild-type CYLD, (S/A)CYLD and (H/N)CYLD, and NF-κB activation was analysed by a luciferase assay after stimulation with TNF-α or IL-1β (10 ng ml⁻¹ for 4 h). Results are representative of at least two experiments done in duplicate; bars indicate the s.d. in each series. Relative expression of the CYLD constructs is shown below the luciferase data. **b**, CYLD-mediated inhibition of NF-κB activation by members of the TNFR and IL-1 receptor families. Analysis was done as in **a**, except that NF-κB was activated by overexpressing the indicated receptors. Relative stimulation is shown; numbers indicate the absolute fold activation. **c**, CYLD-mediated inhibition of NF-κB activation by signalling intermediates. Analysis was done as in **a**. **d**, (R/X)CYLD does not inhibit NF-κB activation. Analysis was done as in **a**, using TRAF2 as activator of NF-κB.

(H/N)CYLD (Fig. 4c and data not shown). This indicates that the overexpressed CYLD is not acting by directly inhibiting the I κ B α ubiquitination step (see also Fig. 3f), the subsequent I κ B degradation step, or probably the final step of IKK activation, but rather functions by inhibiting an upstream event that is required for this activation. Indeed, although it effectively blocked degradation of I κ B induced by TNF treatment, CYLD had no effect on the phosphorylation of I κ B nor on its degradation in response to IKK2(E/E) expression (Supplementary Fig. 4). (R/X)CYLD did not seem to inhibit NF- κ B activation (Fig. 4d), further showing that the negative regulation of NF- κ B activation by CYLD is dependent on its catalytic activity.

To confirm the negative role of CYLD in TRAF-dependent NF- κ B activation, we tried to reduce endogenous CYLD expression through RNA-mediated (RNAi) interference¹⁸ using a pSUPER expression vector (pSUPER-CYLD)¹⁹. Expression of exogenous full-length CYLD, but not of a CYLD truncation mutant lacking the short interfering RNA (siRNA) target site (Δ N), was markedly decreased by expression of this vector (Fig. 5a), showing its efficiency and specificity in reducing CYLD expression. When transfected with pSUPER-CYLD, HEK 293T cells showed a roughly twofold increase in the activation of NF- κ B induced by small amounts of TRAF2. By contrast, the same amount of NF- κ B activation induced by Tax was not changed by expression of the CYLD siRNA, indicating that the effect was specific (Fig. 5b).

Our study shows that the tumour suppressor CYLD is a member of the UBP class of deubiquitinating enzymes that preferentially cleaves non-K48-Ub-conjugated moieties and negatively regulates NF- κ B activation. While this manuscript was in preparation, CYLD was shown to possess deubiquitinating activity on the basis of experiments using active-site-directed inhibitors of deubiquitinating enzymes²⁰. Our data indicate that TRAF2 and NEMO are *in vivo* targets of CYLD. The association of CYLD with NEMO and TRAF2, and probably also cytoskeletal associations imposed on CYLD by its CAP-Gly motifs⁴, may allow it to specifically affect other target proteins. Despite our still limited knowledge of the identity of the CYLD substrates, our findings clearly indicate that this molecule controls the ubiquitination of signalling proteins that function upstream of IKKs, thereby identifying an additional stage of regulation in the NF- κ B pathway. CYLD dysfunction at this stage may result in excessive activation of NF- κ B and trigger cell transformation through increased resistance to apoptosis or perturbation of the cell cycle²¹. □

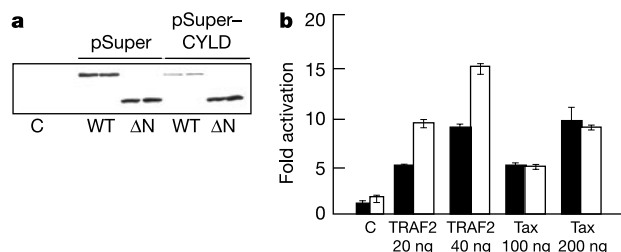


Figure 5 CYLD is a negative regulator of the TRAF2 and NF- κ B signalling pathway. **a**, CYLD siRNA efficiently blocks expression of transfected CYLD. HEK 293T cells were transiently transfected with either empty pcDNA3 (C) or expression vectors encoding Flag-tagged wild-type CYLD or an N-terminally truncated mutant of CYLD (Δ N; residues 1–393 deleted). CYLD expression was analysed by immunoblotting with a monoclonal antibody against Flag. **b**, Expression of CYLD siRNA potentiates TRAF-dependent activation of NF- κ B. HEK 293T cells were co-transfected in duplicate with the indicated amounts of a TRAF2 or Tax expression construct or empty vector (C), plus the I κ B-Luc reporter and either pSUPER-CYLD (open bars) or control pSUPER (filled bars) plasmids. Analysis was done as in Fig. 4.

Methods

Yeast two-hybrid screening

A complementary DNA corresponding to the C-terminal part of the human NEMO open reading frame (ORF; nucleotides 284–1260) was cloned into the GAL4 DNA-binding domain (DBD) vector pGBT9 (Clontech). The resulting plasmid pGBT9-NEMO was used as bait in a two-hybrid screen of a human B-cell cDNA library (Clontech) in *Saccharomyces cerevisiae* HF7c. Isolation of positive clones and subsequent two-hybrid interaction analyses using a Matchmaker TwoHybrid System Protocol (Clontech) were done according to the manufacturer's instructions. We assessed the binding properties of CYLD and other proteins in the yeast SFY526 reporter strain (Clontech) using the pGBT9-GAL4-DBD and pGAD-GH-GAL4-AD vectors. Deletion constructs for two-hybrid mapping were made by conventional polymerase chain reaction (PCR).

Antibodies and expression constructs

We used antibodies against TRAF2 (H-249, Santa Cruz), TRAF6 (H-274, Santa Cruz), NEMO (PharMingen and ref. 5), the Flag epitope (M2, Sigma), the HA epitope (a home-made hybridoma), β -catenin (Sigma) and I κ B α (Transduction Laboratories).

The wild-type ubiquitin expression vector encoding an octameric tandem fusion of HA-ubiquitin was a gift from M. Treier (EMBL). Both the K48R and K63R HA-ubiquitin point mutants in a bacterial expression vector were a gift from K.-I. Nakayama (Kyushu University). For the mammalian expression experiments, we transferred a single copy of each of the mutant cDNAs into the pcDNA3 vector by using conventional cloning techniques.

A cDNA comprising the complete ORF of CYLD was obtained by joining the original two-hybrid clone (clone 10) and a partially overlapping 5'-extending clone obtained from the IMAGE consortium. The resulting full-length cDNA was used to amplify by PCR the predicted CYLD ORF, which was subsequently transferred into a pcDNA3 mammalian expression vector (Invitrogen) that had been modified to include in-frame amino-terminal AU1 and C-terminal Flag epitopes.

We prepared all deletions and point mutants by conventional PCR and cloning techniques. The CYLD R936X deletion construct was derived directly from the C-terminally tagged plasmid pcCYLDwt-Flag and thus contained no epitope tag at its C terminus. For this reason, the activity of this construct was always compared with that of untagged wild-type CYLD.

Cell culture and transfection

HEK 293T cells were transfected as described²². We carried out immunoblot analysis, immunoprecipitations and luciferase assays as described⁵.

In vitro deubiquitination assay

Both the enzyme (CYLD-Flag) and the substrate (Flag-TRAF6) were expressed separately in HEK 293T cells in 10-cm dishes. After immunoprecipitation with antibodies against Flag, CYLD was eluted by incubation with Flag peptide (Sigma) and added to Flag-TRAF6 beads. The deubiquitination reaction was carried out in 100 μ l of deubiquitination buffer (20 mM Tris-HCl (pH 7.5), 5 mM MgCl₂ and 2 mM dithiothreitol) at 30 °C for 1 h. We resolved the reaction products by conventional SDS-PAGE and analysed them by immunoblotting.

CYLD siRNA expression vector

The RNAi procedure was done as described¹⁹. In brief, a double-stranded oligonucleotide was designed to contain a sequence derived from the 5' end of human CYLD ORF (nucleotides 232–250) in forward and reverse orientation separated by a 9-base-pair spacer region (ttcaagaga) to allow formation of the hairpin structure in the expressed oligo-RNA: sense strand, 5'-gatccccCTCATGCAGTTCCTCTTgttcaagagaCAAAGA GAACTGCATGAGGtttttgaaa; antisense strand, 5'-agcttttccaaaacCTCATGCAGTTCCTCTTgtctcttgaaCAAAGAGAACTGCATGAGGggg. The resulting double-stranded oligonucleotide was cloned into the BglII and HindIII sites of the pSUPER vector for expression under the control of the H1 RNA promoter.

Received 17 March; accepted 20 May 2003; doi:10.1038/nature01802.

- Karin, M. & Ben-Neriah, B. Phosphorylation meets ubiquitination: the control of NF- κ B activity. *Annu. Rev. Immunol.* **18**, 621–663 (2000).
- Li, Q. & Verma, I. M. NF- κ B regulation in the immune system. *Nature Rev. Immunol.* **2**, 725–734 (2002).
- Israel, A. The IKK complex: an integrator of all signals that activate NF- κ B? *Trends Cell. Biol.* **10**, 129–133 (2000).
- Bignell, G. R. *et al.* Identification of the familial cylindromatosis tumour-suppressor gene. *Nature Genet.* **25**, 160–165 (2000).
- Yamaoka, S. *et al.* Complementation cloning of NEMO, a component of the I κ B kinase complex essential for NF- κ B activation. *Cell* **93**, 1231–1240 (1998).
- Rothwarf, D. M., Zandi, E., Natoli, G. & Karin, M. IKK- γ is an essential regulatory subunit of the I κ B kinase complex. *Nature* **395**, 297–300 (1998).
- Courtois, G., Smahi, A. & Israël, A. NEMO/IKK γ : linking NF- κ B to human disease. *Trends Mol. Med.* **7**, 427–430 (2001).
- Ye, H., Park, Y. C., Kreishman, M., Kieff, E. & Wu, H. The structural basis for the recognition of diverse receptor sequences by TRAF2. *Mol. Cell* **4**, 321–330 (1999).
- Wilkinson, K. D. Regulation of ubiquitin-dependent processes by deubiquitinating enzymes. *FASEB J.* **11**, 1245–1256 (1997).
- D'Andrea, A. & Pellman, D. Deubiquitinating enzymes: a new class of biological regulators. *Crit. Rev. Biochem. Mol. Biol.* **33**, 337–352 (1998).
- Deng, L. *et al.* Activation of the I κ B kinase complex by TRAF6 requires a dimeric ubiquitin-conjugating enzyme complex and a unique polyubiquitin chain. *Cell* **103**, 351–361 (2000).

12. Wang, C. *et al.* TAK1 is a ubiquitin-dependent kinase of MKK and IKK. *Nature* **412**, 346–351 (2001).
13. Shi, C. S. & Kehrl, J. H. TNF-induced GSK-3 and SAPK activation depends upon the E2/E3 complex Ubc13-Uev1A/TRAFA2. *J. Biol. Chem.* **278**, 15429–15434 (2003).
14. Maniatis, T. A ubiquitin ligase complex essential for the NF- κ B/Wnt/Wingless, and Hedgehog signaling pathways. *Genes Dev.* **13**, 505–510 (1999).
15. Pomerantz, J. L. & Baltimore, D. Two pathways to NF- κ B. *Mol. Cell* **10**, 693–695 (2002).
16. Chu, Z. L., DiDonato, J. A., Hawiger, J. & Ballard, D. W. The tax oncoprotein of human T-cell leukemia virus type 1 associates with and persistently activates I κ B kinases containing IKK α and IKK β . *J. Biol. Chem.* **273**, 15891–15894 (1998).
17. Mercurio, F. *et al.* IKK-1 and IKK-2: cytokine-activated I κ B kinases essential for NF- κ B activation. *Science* **278**, 860–866 (1997).
18. McManus, M. T. & Sharp, P. A. Gene silencing in mammals by small interfering RNAs. *Nature Rev. Genet.* **3**, 737–747 (2002).
19. Brummelkamp, T. R., Bernards, R. & Agami, R. A system for stable expression of short interfering RNAs in mammalian cells. *Science* **296**, 550–553 (2002).
20. Borodovsky, A. *et al.* Chemistry-based functional proteomics reveals novel members of the deubiquitinating enzyme family. *Chem. Biol.* **9**, 1149–1159 (2002).
21. Karin, M., Cao, Y., Greten, F. R. & Li, Z. W. NF- κ B in cancer: from innocent bystander to major culprit. *Nature Rev. Cancer* **2**, 301–310 (2002).
22. Döflinger, R. *et al.* X-linked anhidrotic ectodermal dysplasia with immunodeficiency is caused by impaired NF- κ B signaling. *Nature Genet.* **27**, 77–85 (2001).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank R. Agami for the pSUPER vector and advice; R. Beyaert, C. J. Kirschning, M. Krönke, F. Mercurio, K.-I. Nakayama, M. Rowe and M. Treier for plasmids; T. Goncharov, S. Leu, D. Landstein, M. Pasparakis, F. Agou and S. Yamaoka for discussions and support; and I. Beilis and T. Lopez for technical assistance. This work was supported in part by grants from Inter-Lab Ltd, from Ares Trading SA and from the Alfred and Ann Goldstein Foundation to A.K., G.Ca and D.W., from PTR Pasteur/Necker to G.Co, and from 'La ligue Nationale contre le Cancer' (équipe labellisée) to A.I. A.K. was supported by a postdoctoral fellowship from 'La ligue Nationale contre le Cancer'.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.W. (David.Wallach@weizmann.ac.il) or G.C. (gmcourt@pasteur.fr).

Rationalization of the effects of mutations on peptide and protein aggregation rates

Fabrizio Chitti^{1*}, Massimo Stefani², Niccolò Taddei², Giampietro Rampotti² & Christopher M. Dobson¹

¹Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK

²Dipartimento di Scienze Biochimiche, Università degli Studi di Firenze, Viale Morgagni 50, 50134 Firenze, Italy

* Present address: Dipartimento di Scienze Biochimiche, Università degli Studi di Firenze, Viale Morgagni 50, 50134 Firenze, Italy

In order for any biological system to function effectively, it is essential to avoid the inherent tendency of proteins to aggregate and form potentially harmful deposits^{1–4}. In each of the various pathological conditions associated with protein deposition, such as Alzheimer's and Parkinson's diseases, a specific peptide or protein that is normally soluble is deposited as insoluble aggregates generally referred to as amyloid^{2,3}. It is clear that the aggregation process is generally initiated from partially or completely unfolded forms of the peptides and proteins associated with each disease. Here we show that the intrinsic effects of specific mutations on the rates of aggregation of unfolded polypeptide chains can be correlated to a remarkable extent with changes in simple physicochemical properties such as hydrophobicity, secondary structure propensity and charge. This approach allows the pathogenic effects of mutations associated with known familial forms of protein deposition diseases to

be rationalized, and more generally enables prediction of the effects of mutations on the aggregation propensity of any polypeptide chain.

The ability to form highly organized aggregates such as the fibrils and plaques associated with the amyloid diseases has been suggested to be a generic property of polypeptide chains, and not simply a feature of the small numbers of proteins so far associated with recognized pathological conditions⁵. Nevertheless, the propensity of a given polypeptide chain to aggregate under specific conditions varies dramatically with its composition and sequence. The proposed generic nature⁵ and well established common structural characteristics of aggregates formed from proteins without detectable sequence or structural similarity⁶ have encouraged us to investigate whether the effects of amino acid substitutions on the propensity of proteins to aggregate can be described by relatively simple principles that are of general validity. When a protein that is normally folded starts to aggregate, it does so from at least partially unfolded states that are present during or immediately following biosynthesis, or as the result of cellular stress or proteolytic degradation^{2,3}.

Destabilization of the native state that results in an increased population of partially folded molecules is well established as an important factor in the pathogenic effects of mutations, when the diseases are associated with the deposition of proteins that are globular in their normal functional states². Nevertheless, this factor is not sufficient to explain the pathogenic effects of all mutations associated with such proteins^{7,8}, in part because the mutations also affect the properties of their aggregation-prone unfolded or partially unfolded states. Clarifying such additional factors that modulate aggregation from non-native states is of fundamental importance for a complete understanding of the relationship between mutation and aggregation behaviour for these systems. More importantly, however, the peptides and proteins

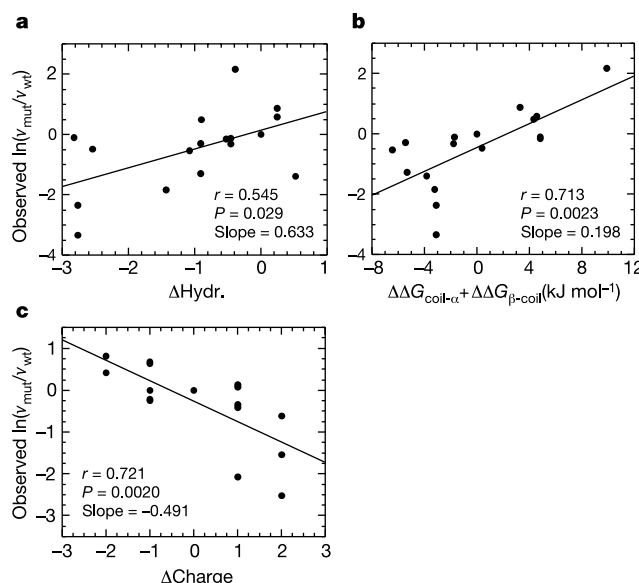


Figure 1 Dependence of the aggregation rate on different physicochemical parameters. Change of the aggregation rate of AcP resulting from mutation plotted against **a**, the predicted change of hydrophobicity, **b**, propensity to convert from an α -helical to a β -sheet conformation, and **c**, charge. All aggregation rate measurements were carried out under conditions in which all the protein variants are substantially unstructured. The mutations reported in **a** and **b**, described previously⁹, involve residues 16–31 and 87–98 and do not involve change of charge. The mutations reported in **c**, also described previously¹⁰, were designed to minimize change of hydrophobicity and secondary structure propensities. Values of $\Delta\text{Hydr.}$, $(\Delta\Delta G_{\text{coil-}\alpha} + \Delta\Delta G_{\text{coil-}\beta})$ and ΔCharge were calculated for each mutation as described in Table 1 legend.

that aggregate in the most prevalent and important neurodegenerative protein deposition disorders, such as Alzheimer's and Parkinson's diseases, as well as some of the non-neuropathic amyloidoses, have been found to adopt predominantly unstructured conformations in their normal biological state (ref. 3 and references therein). For all these systems, aggregation does not need to be preceded by the unfolding of a globular structure, but involves direct self-assembly of the peptide or protein involved from an ensemble of unstructured conformations.

In the light of the importance of aggregation from at least relatively unstructured states of polypeptide chains, the starting point of our approach has been a detailed mutational study of the aggregation process of a small globular protein under conditions in which the native states of all the studied mutational variants are unstable and at most only marginally populated. Human muscle acylphosphatase (AcP) is highly amenable to such studies, as its aggregation behaviour is well defined and extremely reproduc-

ible^{9,10}. In particular, the rate of aggregation of AcP from an ensemble of denatured conformations can readily be followed by using a variety of spectroscopic probes, and has now been measured for over 50 mutational variants of this protein^{9,10}. By using conditions under which the protein is denatured, changes in aggregation rates can be attributed directly to the intrinsic effects of the amino acid substitutions on the process of self-assembly, without the complications of additional contributions arising from the destabilization of the native state. Many of the mutations examined, particularly those involving residues in two key regions, 16–31 and 87–98, were found to perturb the aggregation rate of AcP very significantly^{9,10}. In order to analyse the results, we have characterized all the amino acid substitutions in terms of their effects on three parameters that we anticipated to be important in influencing aggregation rates. These are the change in the hydrophobicity of the polypeptide chain resulting from mutation ($\Delta\text{Hydr.}$), in the propensity to convert from α -helical to β -sheet structure ($\Delta\Delta G_{\text{coil-}\alpha} + \Delta\Delta G_{\beta\text{-coil}}$) and in

Table 1 Changes resulting from single-point mutations of unstructured peptides or natively unfolded proteins

Mutation	$\Delta\text{Hydr. (kcal mol}^{-1}\text{)}^*$	$\Delta\Delta G_{\beta\text{-coil}} \text{ (kJ mol}^{-1}\text{)}^\dagger$	$\Delta\Delta G_{\text{coil-}\alpha} \text{ (kJ mol}^{-1}\text{)}^\ddagger$	$\Delta\text{Charge}^\S$	Calculated $\ln(\nu_{\text{mut}}/\nu_{\text{wt}})^\parallel$	Observed $\ln(\nu_{\text{mut}}/\nu_{\text{wt}})^\P$	ref. #
Amylin							
N22A	2.30	-0.95	-3.36	0	0.60	0.70 ± 0.60	13
F23A	-1.88	-4.64	-3.90	0	-2.88	-2.65 ± 0.50	13
G24A	0.39	1.77	-2.84	0	0.04	-0.05 ± 0.30	13
I26A	-1.43	-5.05	-0.32	0	-1.97	-2.40 ± 0.50	13
L27A	-1.43	-2.05	0.36	0	-1.24	-0.95 ± 0.50	13
S20G	1.24	-4.09	0.00	0	-0.03	1.00 ± 0.40	14
Prion peptides							
H111A	3.26	-1.36	-3.21	-1	1.65	0.60 ± 0.70	15
H111K	0.10	0.41	-1.72	0	-0.20	-0.25 ± 0.10	15
A117V☆	0.91	4.63	2.37	0	1.96	1.50 ± 0.70	15
V210I☆	0.52	0.41	-0.97	0	0.22	0.85 ± 0.10	16
α-Synuclein							
A53T☆	-1.39	5.59	2.83	0	0.79	1.20 ± 0.40	17
A76E	-3.30	1.64	0.00	1	-2.25	-2.70 ± 0.70	12
A76R	-4.34	1.64	0.64	-1	-1.80	-0.90 ± 0.40	12
Amyloid β-peptide**							
A21G☆	-0.39	-1.77	3.27	0	0.05	-0.07 ± 0.30	18
E22K☆	0.14	0.14	-1.72	-2	0.76	0.90 ± 0.40	19
E22Q☆	1.61	0.14	0.00	-1	1.54	2.90 ± 0.80	18,19
E22G☆	2.91	-3.41	4.30	-1	2.51	2.05 ± 0.30	20
D23N☆	1.90	4.36	-1.72	-1	2.22	3.95 ± 0.80	18
F19T	-3.27	0.95	-1.76	0	-2.23	-2.50 ± 0.20	21
Tau							
R5L☆	5.77	0.40	-1.00	-1	4.02	3.05 ± 1.50	22
G272V☆	1.30	6.41	-1.71	0	1.75	1.05 ± 0.50	23,24
R406W☆	6.08	1.50	0.00	-1	4.64	1.25 ± 0.80	23–25
Y310W	0.66	-1.77	0.00	0	0.07	0.05 ± 0.05	26
Leucine-rich repeat							
D24N	1.90	4.36	-3.43	-1	1.88	2.10 ± 1.00	27
D24Q	2.51	5.18	-3.10	-1	2.49	1.25 ± 0.70	27
Model peptide							
D6E	0.90	5.04	-2.27	0	1.12	0.40 ± 0.30	28
D6N	1.90	4.36	0.00	1	1.57	0.50 ± 0.30	28

* Calculated using $\Delta\text{Hydr.} = \text{Hydr.}_{\text{wt}} - \text{Hydr.}_{\text{mut}}$, where $\Delta\text{Hydr.}$ is the change of hydrophobicity resulting from mutation, and Hydr._{wt} and $\text{Hydr.}_{\text{mut}}$ are the hydrophobicity values of the wild-type and mutant residues, respectively (the values of hydrophobicity for all 20 amino acids are listed in the table provided in Supplementary Information 1).

† Calculated using $\Delta\Delta G_{\beta\text{-coil}} = 13.64 (P_{\beta}^{\text{wt}} - P_{\beta}^{\text{mut}})$, where $\Delta\Delta G_{\beta\text{-coil}}$ is the change of free energy change for the transition random coil $\rightarrow$ β -sheet resulting from mutation, P_{β}^{wt} and P_{β}^{mut} are the β -sheet propensities of the wild-type and mutant residue, respectively (the values of β -sheet propensity for all 20 amino acids are listed in the table provided in Supplementary Information 1), and 13.64 is the conversion constant from the normalized scale to units of kJ mol^{-1} .

‡ Calculated using $\Delta\Delta G_{\text{coil-}\alpha} = RT \ln(P_{\alpha}^{\text{wt}}/P_{\alpha}^{\text{mut}})$, where $\Delta\Delta G_{\text{coil-}\alpha}$ is the predicted change of free energy change for the transition α -helix $\rightarrow$ random coil resulting from mutation, P_{α}^{wt} and P_{α}^{mut} are the predicted α -helical propensities (helix percentages) of the wild-type and mutated sequences at the site of mutation, respectively (calculated using the AGADIR algorithm at www.embl-heidelberg.de/Services/serrano/agadir/agadir-start.html); $R=0.008314 \text{ kJ mol}^{-1} \text{ K}^{-1}$.

§ Calculated using $\Delta\text{Charge} = |\text{Charge}_{\text{mut}}| - |\text{Charge}_{\text{wt}}|$, where ΔCharge is the change of charge resulting from the mutation, and $|\text{Charge}_{\text{mut}}|$ and $|\text{Charge}_{\text{wt}}|$ are the absolute values of charge for the mutated and wild-type sequences, respectively. The operator of 'absolute value' is introduced so that a negative value of ΔCharge results from the equation when the mutation causes the entire protein or peptide to approach neutrality, regardless of the initial sign of charge of the wild-type protein. A positive value of ΔCharge is obtained when the mutation causes the entire protein sequence to deviate further from neutrality.

|| Predicted change of aggregation rate resulting from mutation, calculated using equation (1) (see text for details).

¶ Observed change of aggregation rate resulting from mutation (obtained from published experimental data as referenced in the last column of the table). Salt concentrations were within physiological values in the various experiments reported in the table.

References from which the values of observed $\ln(\nu_{\text{mut}}/\nu_{\text{wt}})$ are taken.

☆ Natural mutations associated with familial diseases.

** Previously published mutations involving substitutions to proline could not be included in this analysis because of the difficulty in calculating the β -sheet propensities of proline residues from experimental data²⁹. Other mutations were also not considered because only qualitative measurements of the effect of such mutations on the aggregation rate were reported. Good agreement is, however, found between our predictions and the qualitative data determined experimentally.

the overall charge (ΔCharge). These changes can be quantified using published values of these parameters for all the amino acid residues (reported in Supplementary Information) and the functions given in Table 1 legend.

The change of aggregation rate of AcP upon mutation ($\ln(\nu_{\text{mut}}/\nu_{\text{wt}})$; see Methods for details) was found to correlate significantly with each of these parameters (Fig. 1a–c). In each of the three plots, however, the data points are considerably scattered around the lines representing the best fits to linear functions. This scatter can readily be attributed to the fact that only a single parameter is considered in each case, to the limitations in the presently available data for predicting accurately changes in the hydrophobicity and secondary structure propensities, and to the varying relative importances of the different sites of mutation in the aggregation process. Despite the scatter present in each plot, however, all three correlations are statistically significant, allowing the mean dependencies of $\ln(\nu_{\text{mut}}/\nu_{\text{wt}})$ on these three parameters to be calculated. These mean values are represented by the slopes of the lines of best fit resulting from the analysis.

On the assumption that the three factors identified as influencing aggregation are independent of each other and additive, we used a combined function to predict the effect of a mutation on aggregation rate:

$$\ln(\nu_{\text{mut}}/\nu_{\text{wt}}) = A\Delta\text{Hydr.} + B(\Delta\Delta G_{\text{coil}-\alpha} + \Delta\Delta G_{\beta\text{-coil}}) + C\Delta\text{Charge} \quad (1)$$

Our best present estimates of the A , B and C factors are the slopes of the three plots reported in Fig. 1 (that is, the dependencies of $\ln(\nu_{\text{mut}}/\nu_{\text{wt}})$ on the three parameters). These are 0.633, 0.198 and -0.491 , for A , B and C , respectively.

This function was then used to calculate the changes of the aggregation rate resulting from mutations in a range of other peptides and proteins, particularly those associated with neurodegenerative protein deposition diseases. In order to test this approach, we considered mutations that complied with the following criteria: (1) the experimental value of $\ln(\nu_{\text{mut}}/\nu_{\text{wt}})$ is directly available or can be determined from data reported in the literature, and (2) the effect of the mutation is reported under conditions in which the polypeptide chain is unstructured. The mutations considered in this analysis, together with the corresponding experimental and calculated values of $\ln(\nu_{\text{mut}}/\nu_{\text{wt}})$, are listed in Table 1. These mutations include both physiologically relevant ones associ-

ated with familial forms of protein deposition diseases (identified in Table 1 with the symbol ☆) and other substitutions that have been investigated to address the importance of particular residues in the aggregation of specific systems. As many peptides or proteins that undergo aggregation in protein deposition diseases are substantially unstructured under physiological conditions, most of the experimental data we could find, and are considered here, were obtained under conditions close to physiological. The various peptide and protein systems reported in Table 1 were studied at different concentrations, but by taking the ratio of ν_{mut} to ν_{wt} the data are normalized in such a manner as to allow for this effect and indeed for the variability in solution conditions.

Figure 2a shows the resulting plot of calculated versus experimental values of $\ln(\nu_{\text{mut}}/\nu_{\text{wt}})$ for the 27 mutations that we identified that satisfied these criteria. The correlation ($r = 0.85$, $P < 0.0001$) is highly significant, and the value of the slope (0.94) is close to 1.0, indicating that there is remarkably good agreement between the calculated and experimental effects of mutations on the aggregation rates. The observed changes of aggregation rate resulting from mutation span a range of nearly 1,000, from 15 times slower to over 50 times faster than the corresponding wild-type polypeptide chains (Fig. 2a and Table 1). Figure 2b shows the calculated versus experimental values of $\ln(\nu_{\text{mut}}/\nu_{\text{wt}})$ for all the mutations of AcP within the two key regions of the sequence 16–31 and 87–98. The observed correlation is again highly significant ($r = 0.756$ and $P < 0.0001$) and the slope is close to 1.0. It is remarkable that the correlation obtained from mutations of proteins or peptides other than AcP (Fig. 2a) is as significant as that obtained with data from the AcP mutations (Fig. 2b), that is, with data used for the derivation of equation (1).

Examples where close agreement is found between theoretical and experimental values include mutations associated with hereditary spongiform encephalopathies, early-onset Parkinson's disease, early-onset Alzheimer's disease and hereditary cerebral haemorrhage with amyloidosis (Table 1). The effects of mutations that inhibit aggregation, and hence do not cause disease, are also predicted as effectively as those that favour it. Examples include the F19T mutation of the amyloid β -peptide, the A76E mutation of α -synuclein and the L27A mutation of amylin (Table 1). For some mutations, such as the R406W mutation of tau and the V210I mutation of a prion peptide, the agreement is quantitatively less good. However, even in these cases our combined function is generally able to predict qualitatively whether the mutations have an accelerating, decelerating, or only a small effect on the corresponding polypeptides (Table 1).

Interestingly, application of equation (1) reveals that many mutations are predicted not to increase but to reduce very significantly the rates of aggregation of the peptides and proteins that we investigated. It therefore appears that the sequences of natural proteins have not been fully optimized during evolution to resist aggregation, but simply to be sufficiently resistant to aggregation to remain soluble during the normal reproductive life spans¹¹. This finding gives rise to the intriguing possibility that there could be a range of mutations within the human population that reduce as well as increase the susceptibility to conditions such as Alzheimer's disease. It will be extremely interesting to test such a prediction, for example by studies of elderly people not afflicted with such illness.

The correlation found in Fig. 2a between the theoretical and experimental effects of mutations on aggregation rates is particularly striking if one considers the heterogeneous groups of peptide and protein systems used in the analysis. Moreover, it is clear that the aggregation rate of a polypeptide chain is more sensitive to mutations occurring within some, often relatively restricted, regions of the sequence^{9,12}. The explanation of this latter point is most probably that data reported in the literature are very likely to be those within such regions, as they involve either disease-related

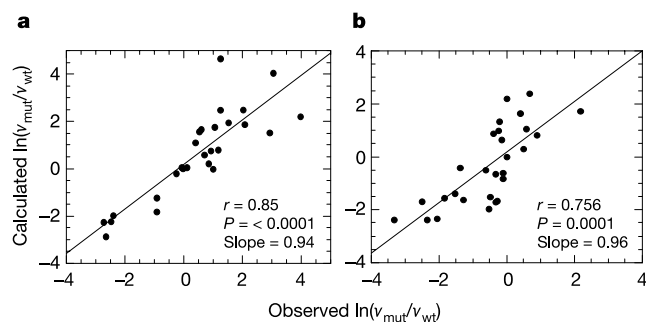


Figure 2 Calculated versus observed changes in aggregation rate upon mutation. **a**, Data relating to mutations of short peptides or natively unfolded proteins, including amylin, amyloid β -peptide, tau and α -synuclein. The 27 mutations shown in the plot and their calculated and experimental $\ln(\nu_{\text{mut}}/\nu_{\text{wt}})$ values are those listed in Table 1. **b**, Data for the mutations of AcP that are within the two regions of the sequence that have been found to be particularly important for aggregation (encompassing residues 16–31 and 87–98)⁹. All measurements of aggregation rates were carried out under conditions in which all the protein variants are substantially unstructured. In both the analyses (**a** and **b**) the calculated $\ln(\nu_{\text{mut}}/\nu_{\text{wt}})$ values were obtained using equation (1).

mutations or substitutions aimed at identifying such residues in specific experimental studies. Clearly it is important when predicting the effects of mutations either to ensure that they are within such regions or to accept that the predictions could be somewhat higher than the real values.

With these caveats, however, the finding that the effects of mutations on the aggregation rates of these different proteins are closely similar strongly suggests that they result from perturbations to fundamental physicochemical properties of the polymer chains rather than from alterations of specific interactions involving the side chains within the resulting aggregates. This strongly supports the conclusion that there are common principles in the fundamental determinants of aggregation of polypeptide chains despite evident differences in detail^{5,30}. Moreover, most of the disease-associated mutations analysed here clearly increase the rate of aggregation (Table 1). This observation indicates that these mutations are likely to be pathogenic primarily as a result of an accelerated rate of aggregation originating from the perturbation of the fundamental parameters that we are now able to identify and describe in a quantitative manner.

The approach described here enables us to begin to identify the factors that affect the intrinsic propensities of polypeptide chains to aggregate. As additional data become available it will undoubtedly be possible to refine further this approach, and also to consider additional parameters that influence the aggregation propensities of peptides and proteins. Even at the present level of sophistication, however, we have shown that a simple function has the ability to rationalize and to predict to a remarkable degree the effects of mutations on the aggregation rates of a wide range of peptides and proteins. As well as giving important insights into the origins of protein misfolding diseases, its use should help substantially in efforts to modify rationally the aggregational properties of proteins or peptides that at present can seriously limit their utilization in research programmes, for medical applications or for biotechnological purposes. □

Methods

The rates of aggregation of the various AcP variants were measured in 25% trifluoroethanol from the time courses of thioflavine T fluorescence, as previously described⁹. The change of aggregation rate as a result of a mutation is expressed in all cases as the natural logarithm of the ratio of the aggregation rate constants of the mutant and wild-type protein ($\ln(\nu_{\text{mut}}/\nu_{\text{wt}})$).

The experimental values of $\ln(\nu_{\text{mut}}/\nu_{\text{wt}})$ for mutations of proteins other than AcP were obtained from published data (see last column of Table 1 for references of each mutation). Only data reporting the effect of single-point mutations on the aggregation of proteins and peptides under conditions in which they are unstructured were considered. Data were considered regardless of the experimental techniques used by the different authors to probe aggregation, provided that a quantitative analysis was carried out. When time or rate constants were not explicitly reported, the plots describing the kinetic profiles of aggregation were scanned and computer-analysed to obtain rate constant values. When lag and growth phases were evident in the kinetic profiles of aggregation only the growth phase was considered. When data at fixed periods of time were reported (for example by means of bar graphs), the experimental $\ln(\nu_{\text{mut}}/\nu_{\text{wt}})$ value was obtained from the ratio of the aggregation parameters obtained from the mutated and wild-type protein/peptide, before equilibrium was reached.

Mutations involving proline residues were not analysed because of the difficulty in obtaining quantitative estimates of the change of β -sheet propensity as a result of these mutations (see Table 1). Nor were mutations considered when substantial discrepancies in the $\ln(\nu_{\text{mut}}/\nu_{\text{wt}})$ value were reported by different authors (when significant but not substantial discrepancies were present, we considered $\ln(\nu_{\text{mut}}/\nu_{\text{wt}})$ values resulting from averages of the available data). The reader is referred to Supplementary Information for a detailed description of how the experimental values of $\ln(\nu_{\text{mut}}/\nu_{\text{wt}})$ were calculated for all mutations.

Received 24 April; accepted 26 June 2003; doi:10.1038/nature01891.

1. Kopito, R. R. Aggregates, inclusion bodies and protein aggregation. *Trends Cell Biol.* **10**, 524–530 (2000).

- Kelly, J. W. The alternative conformations of amyloidogenic proteins and their multi-step assembly pathways. *Curr. Opin. Struct. Biol.* **8**, 101–106 (1998).
- Rochet, J. C. & Lansbury, P. T. Jr Amyloid fibrillogenesis: Themes and variations. *Curr. Opin. Struct. Biol.* **10**, 60–68 (2000).
- Carri6, M. M. & Villaverde, A. Construction and deconstruction of bacterial inclusion bodies. *J. Biotech.* **96**, 3–12 (2002).
- Dobson, C. M. Protein misfolding, evolution and disease. *Trends Biochem. Sci.* **24**, 329–332 (1999).
- Sunde, M. & Blake, C. The structure of amyloid fibrils by electron microscopy and X-ray diffraction. *Adv. Protein Chem.* **50**, 123–159 (1997).
- Liemann, S. & Glockshuber, R. Influence of amino acid substitutions related to inherited human prion diseases on the thermodynamic stability of the cellular prion protein. *Biochemistry* **38**, 3258–3267 (1999).
- Hammarstrom, P., Wiseman, R. L., Powers, E. T. & Kelly, J. W. Prevention of transthyretin amyloid disease by changing protein misfolding energetics. *Science* **299**, 713–716 (2003).
- Chiti, F. *et al.* Kinetic partitioning of protein folding and aggregation. *Nature Struct. Biol.* **9**, 137–143 (2002).
- Chiti, F. *et al.* Studies of the aggregation of mutant proteins *in vitro* provide insights into the genetics of amyloid diseases. *Proc. Natl Acad. Sci. USA* **99**, 16419–16426 (2002).
- Dobson, C. M. Getting out of shape. *Nature* **418**, 729–730 (2002).
- Glasson, B. L., Murray, I. V., Trojanowski, J. Q. & Lee, V. M. A hydrophobic stretch of 12 amino acid residues in the middle of α -synuclein is essential for filament assembly. *J. Biol. Chem.* **276**, 2380–2386 (2001).
- Azriel, R. & Gazit, E. Analysis of the minimal amyloid-forming fragment of the islet amyloid polypeptide. An experimental support for the key role of the phenylalanine residue in amyloid formation. *J. Biol. Chem.* **276**, 34156–34161 (2001).
- Sakagashira, S. *et al.* S20G mutant amylin exhibits increased *in vitro* amyloidogenicity and increased intracellular cytotoxicity compared to wild-type amylin. *Am. J. Pathol.* **157**, 2101–2109 (2000).
- Salmona, M. *et al.* Molecular determinants of the physicochemical properties of a critical prion protein region comprising residues 106–126. *Biochem. J.* **342**, 207–214 (1999).
- Thompson, A. J., Barnham, K. J., Norton, R. S. & Barrow, C. J. The Val-210-Ile pathogenic Creutzfeldt-Jakob disease mutation increases both the helical and aggregation propensities of a sequence corresponding to helix-3 of PrP(C). *Biochim. Biophys. Acta* **1544**, 242–254 (2001).
- Conway, K. A. *et al.* Acceleration of oligomerization, not fibrillization, is a shared property of both α -synuclein mutations linked to early-onset Parkinson's disease: Implications for pathogenesis and therapy. *Proc. Natl Acad. Sci. USA* **97**, 571–576 (2000).
- Van Nostrand, W. E., Melchor, J. P., Cho, H. S., Greenberg, S. M. & Rebeck, G. W. Pathogenic effects of D23N Iowa mutant amyloid β -protein. *J. Biol. Chem.* **276**, 32860–32866 (2001).
- Miravalle, L. *et al.* Substitutions at codon 22 of Alzheimer's β -peptide induce diverse conformational changes and apoptotic effects in human cerebral endothelial cells. *J. Biol. Chem.* **275**, 27110–27116 (2000).
- Nilsberth, C. *et al.* The 'Arctic' APP mutation (E693G) causes Alzheimer's disease by enhanced A β protofibril formation. *Nature Neurosci.* **4**, 887–893 (2001).
- Esler, W. P. *et al.* Point substitution in the central hydrophobic cluster of a human β -amyloid congener disrupts peptide folding and abolishes plaque competence. *Biochemistry* **35**, 13914–13921 (1996).
- Gamblin, T. C., Berry, R. W. & Binder, L. I. Tau polymerization: Role of the amino terminus. *Biochemistry* **42**, 2252–2257 (2003).
- Barghorn, S. *et al.* Structure, microtubule interactions, and paired helical filament aggregation by tau mutants of frontotemporal dementias. *Biochemistry* **39**, 11714–11721 (2000).
- Gamblin, T. C. *et al.* *In vitro* polymerization of tau protein monitored by laser light scattering: Method and application to the study of FTDP-17 mutants. *Biochemistry* **39**, 6136–6144 (2000).
- Nacharaju, P. *et al.* Accelerated filament formation from tau protein with specific FTDP-17 missense mutations. *FEBS Lett.* **447**, 195–199 (1999).
- Li, L., Von Bergen, M., Mandelkow, E. M. & Mandelkow, E. Structure, stability, and aggregation of paired helical filaments from tau protein and FTDP-17 mutants probed by tryptophan scanning mutagenesis. *J. Biol. Chem.* **277**, 41390–41400 (2002).
- Symmons, M. F., Buchanan, S. G., Clarke, D. T., Jones, G. & Gay, N. J. X-ray diffraction and far-UV CD studies of filaments formed by a leucine-rich repeat peptide: Structural similarity to the amyloid fibrils of prions and Alzheimer's disease β -protein. *FEBS Lett.* **412**, 397–403 (1997).
- Orpiszewski, J. & Benson, M. D. Induction of beta-sheet structure in amyloidogenic peptides by neutralization of aspartate: A model for amyloid nucleation. *J. Mol. Biol.* **289**, 413–428 (1999).
- Street, A. G. & Mayo, S. L. Intrinsic β -sheet propensities result from van der Waals interactions between side chains and the local backbone. *Proc. Natl Acad. Sci. USA* **96**, 9074–9076 (1999).
- Bucciantini, M. *et al.* Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature* **416**, 507–511 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

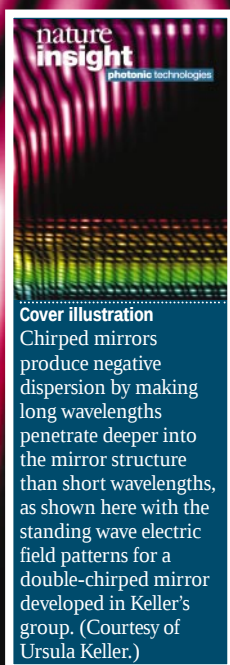
Acknowledgements We thank M. Calamai, J. Zurdo and M. Vendruscolo for critical reading of the manuscript. The Dipartimento di Scienze Biochimiche in Florence is supported by the Italian MIUR (Progetto "Genetica Molecolare" and Progetto FIRB "Folding di proteine: l'altra metà del codice genetico"). We are grateful for support from the Wellcome Trust (F.C. and C.M.D.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.M.D. (cmd44@cam.ac.uk).

nature insight

photonic technologies



Photonics—it's light, but not as we know it. Over the past couple of decades, new techniques have moulded the flow of light in useful and unusual ways. It could be argued that the field of 'photonics' was launched with the discovery of the laser in the 1960; with the advent of this pure, stable and bright light source, new opportunities arose for making use of light, of which perhaps the most notable is the adoption of light as a carrier of information. The now widespread implementation of fibre-optic telecommunication networks is a vivid demonstration of just how fast the technology has developed.

But photonic scientists are not content with using light as simply an information carrier; they are also exploring ways to exploit photons—the basic unit of light—to process and store that information. For instance, an optical circuit can, in principle, be constructed from a new type of optical material, called photonic crystals, in which photons can be manipulated in a manner similar to electrons in a semiconductor. Photons can travel much faster than electrons, which makes the prospect of processing information photonically tantalizing. Photonic chips that outperform their electronic counterparts are a long way from practical realization, but the idea in itself is a source of inspiration for scientists and engineers.

And yet there is more to photonics than imitating electronics. New applications are expected to result from studies where the flow of light is carefully controlled on small lengthscales or fast timescales. There is also a lot still to be learned about the interaction between light and matter. For example, how light is confined within small resonant cavities, how light is 'sculpted' to control the forces that it exerts on particles and how light is coupled to metallic structures smaller than its wavelength. Some of the research avenues that are currently being explored will lead to new optical technologies, and the purpose of this Insight is to highlight some of the most intriguing and promising directions.

Liesbeth Venema Senior Editor

review articles:

- 810 A revolution in optical manipulation**
D. G. Grier
-
- 817 Discretizing light behaviour in linear and nonlinear waveguide lattices**
D. M. Christodoulides, F. Lederer & Y. Silberberg
-
- 824 Surface plasmon subwavelength optics**
W. L. Barnes, A. Dereux & T. W. Ebbesen
-
- 831 Recent developments in compact ultrafast lasers**
U. Keller
-
- 839 Optical microcavities**
K. Vahala
-
- 847 Photonic crystal fibres**
J. C. Knight
-
- 852 Photonics structures in biology**
P. Vukusic & J. R. Sambles
-

Editor, *Nature*: Philip Campbell
Publisher: Peter Collins
Insights Editor: Karl Ziemelis
Editorial Assistant: Simon Gibson

Production Editor: Maria Hodges
Art Director: Majo Xeridat
Art Editor: Martin Harrison
Layouts: Clifford Saunders

Diagrams: Ann Thomson
Production: Sue Gray
Marketing: Stephanie Liu
Sponsorship: Mark Greene

A revolution in optical manipulation

David G. Grier

Department of Physics, James Franck Institute and Institute for Biophysical Dynamics, The University of Chicago, 5640 S. Ellis Avenue, Chicago, IL 60637, USA (e-mail: grier@elbereth.uchicago.edu)

Optical tweezers use the forces exerted by a strongly focused beam of light to trap and move objects ranging in size from tens of nanometres to tens of micrometres. Since their introduction in 1986, the optical tweezer has become an important tool for research in the fields of biology, physical chemistry and soft condensed matter physics. Recent advances promise to take optical tweezers out of the laboratory and into the mainstream of manufacturing and diagnostics; they may even become consumer products. The next generation of single-beam optical traps offers revolutionary new opportunities for fundamental and applied research.

A new generation of techniques that use the forces exerted by carefully sculpted wavefronts of light offers precisely the level of access and control needed for rapid progress at the frontiers of several branches of science and engineering. In particular, optical forces are ideally suited to manipulating mesoscopic systems, which are characterized by length scales ranging from tens of nanometres to hundreds of micrometres, forces ranging from femtonewtons to nanonewtons and time scales ranging upward from a microsecond. In biology, this range covers many of the inter- and intracellular processes responsible for respiration, reproduction and signalling. In physics and chemistry, it corresponds to the still-puzzling interface between classical and quantum mechanical behaviour, which is made all the more perplexing by the general inapplicability of statistical many-body theory in this realm. Fulfilment of the promise of mesoscopic engineering has been held back by the need for tiny motors to drive micromachines and for robust human-scale interfaces with atomic-scale nanotechnology. Until quite recently, the options for manipulating, analysing and organizing mesoscopically textured matter have been limited. The advent of flexible multifunctional optical traps meets this need.

Many of the most powerful optical manipulation techniques are derived from single-beam optical traps known as optical tweezers (see Fig. 1), which were introduced by Arthur Ashkin, Steven Chu and their coworkers at AT&T Bell Laboratories^{1,2}. An optical tweezer uses forces exerted by a strongly focused beam of light to trap small objects. Although the theory behind optical tweezers is still being developed, the basic principles are straightforward for objects either much smaller than the wavelength of light or much larger. Small objects develop an electric dipole moment in response to the light's electric field, which, generally speaking, is drawn up intensity gradients in the electric field toward the focus. Larger objects act as lenses, refracting the rays of light and redirecting the momentum of their photons. The resulting recoil draws them toward the higher flux of photons near the focus³. This recoil is all but imperceptible for a macroscopic lens but can have a substantial influence on mesoscopic objects.

Optical gradient forces compete with radiation pressure resulting from the momentum absorbed or otherwise transferred from the photons in the beam, which acts like a fire hose to blow particles down the optical axis. Stable trapping requires the axial gradient force to dominate, and is achieved when the beam diverges rapidly enough away from the focal point. For this reason, optical tweezers are

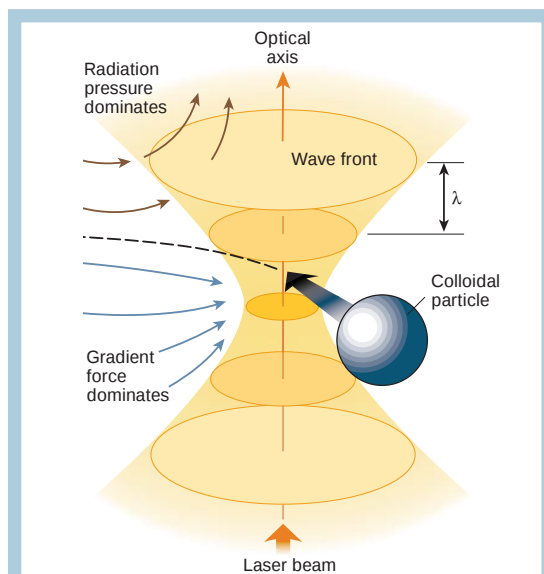


Figure 1 Optical tweezers use a strongly focused beam of light to trap objects. Intensity gradients in the converging beam draw small objects, such as a colloidal particle, toward the focus, whereas the radiation pressure of the beam tends to blow them down the optical axis. Under conditions where the gradient force dominates, a particle can be trapped, in three dimensions, near the focal point.

usually constructed around microscope objective lenses, whose high numerical apertures and well corrected aberrations focus light as tightly as possible.

Optical tweezers can trap objects as small as 5 nm (refs 4,5) and can exert forces exceeding 100 pN (refs 6–8) with resolutions as fine as 100 aN (refs 9–11). This is the ideal range for exerting forces on biological and macromolecular systems and for measuring their responses. Biological and medical applications of optical tweezers have been reviewed extensively^{2,12,13}, and so just a few examples of their uses will be outlined. Optical tweezers have been used to probe the viscoelastic properties of single biopolymers (such as DNA), cell membranes, aggregated protein fibres (such as actin), gels of such fibres in the cytoskeleton, and composite structures (such as chromatin and chromosomes). They have also been used to characterize the forces exerted by molecular motors such as myosin, kinesin, processive enzymes and ribosomes. These measurements have revealed that cells use mechanical forces not only for mobility, motility and chromosome sorting during

reproduction but also for regulating gene transcription, inter- and intracellular signalling and respiration. As a natural extension of these studies, optical tweezers offer great promise for intracellular surgery, for instance, in modifying the chromosomes of living cells¹⁴. On a larger scale, optical tweezers are useful for selecting individual microbes from heterogeneous populations. In addition, their ability to transport and modify cells precisely has led to clinical applications in such areas as *in vitro* fertilization¹⁵.

In the physical sciences, the unique ability of optical tweezers to organize matter non-invasively has led to a burst of activity in the field of classical statistical mechanics, including the first direct measurements of macromolecular interactions in solution¹⁶. Each new round of measurements has led to surprises, including the discovery of anomalous attractions between like-charged colloidal particles¹⁷, oscillatory colloidal interactions mediated by the entropy of smaller entities in solution^{18–21} and hydrodynamic fluctuations that may be interpreted as transient violations of the second law of thermodynamics²².

In all of these cases, and a great many more, fundamental insights emerged from manipulating specially chosen systems at just one or two discrete points. New frontiers of science and engineering would present themselves if optical traps could interrogate more general and more complex systems at many points at once, if they could induce chemical as well as physical transformations and if they could exert torques as well as forces. Recent advances in physical optics reveal that such multifunctional optical traps can be crafted from single beams of light by subtly modifying their wavefronts. The resulting optical micromanipulators provide unprecedented access to the microscopic world.

Manipulating the microscopic world

Figure 2 schematically depicts an optical tweezer system in which a strongly converging objective lens focuses beams of laser light into optical traps. A collimated TEM₀₀ beam passing straight into the input pupil of the lens comes to a focus in the middle of the focal plane of the objective lens, where it forms a trap. Sweeping the angle of incidence translates the trap across the field of view. If the beam diverges, it focuses downstream of the focal plane, whereas if it converges, it focuses upstream.

Translating an optical trap, creating multiple optical traps and converting these into multifunctional optical traps are greatly facilitated by first forming an image of the input pupil of the lens using the telescope in Fig. 2. Any beam passing through the pupil's image, which is centred at point A in Fig. 2, also passes through the actual pupil and forms a trap. Tilting the beam as it passes through the image scans the optical trap. A single rapidly scanned optical tweezer can trap multiple particles by dwelling briefly on each one before moving on to the next^{23,24}. The extent and complexity of such multiparticle patterns are limited by the time required to reposition each of the multiple wandering objects. Scanned optical tweezers, furthermore, are restricted to the focal plane of the lens. Even so, scanned optical tweezers are extremely useful for organizing planar assemblies of colloidal particles²⁵, for testing new ideas in statistical mechanics²⁶ and for measuring macromolecular interactions²⁷.

Placing a diffractive beamsplitter at the pupil's image converts a single input beam into several beams, each of which forms a separate optical trap. Such a beamsplitter can be a computer-generated hologram, and the resulting trapping patterns are known as holographic optical tweezers (HOTs)^{28,29}. To see how this works, consider multiple beams all passing simultaneously through point A on their way to being focused into optical traps. Their superposition creates a distinctive interference pattern centred at point A. Imprinting this pattern onto the wavefronts of a single input laser beam transforms the one beam into the desired fan-out of beams and thus forms the same pattern of optical traps.

The input beam's electric field, $E(\rho) \exp(i\varphi(\rho)) \epsilon$, around point A is characterized by a real-valued amplitude $E(\rho)$ and phase $\varphi(\rho)$, both of which are functions of position transverse to the optical axis

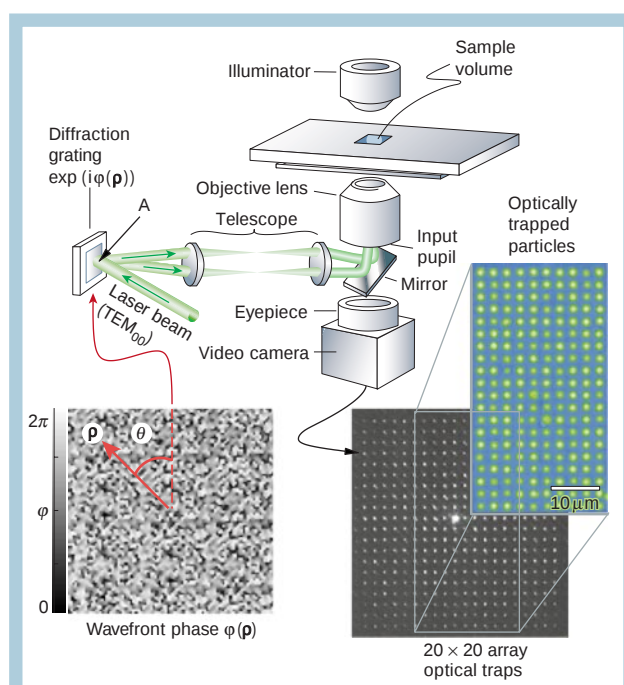


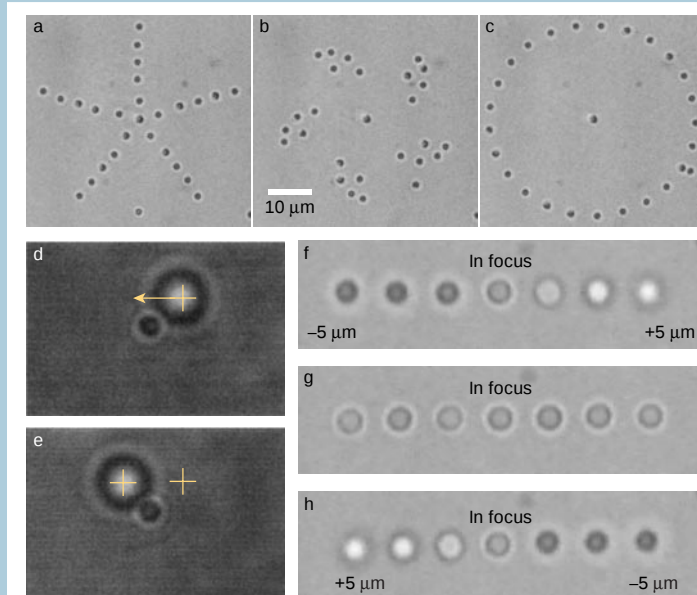
Figure 2 Creation of a large number of optical tweezers by using a computer-generated hologram. Projecting a collimated TEM₀₀ laser beam through the input pupil of a strongly converging lens such as a microscope objective creates a single optical tweezer. The telescope in this implementation creates an image of the objective's input pupil, centred at point A. Multiple beams passing through point A therefore pass into the objective lens to create multiple optical traps. A single TEM₀₀ laser beam can be split into an arbitrary fan-out of beams all emanating from point A by an appropriate computer-designed diffraction grating centred there. The example phase grating $\varphi(\rho)$ creates the 20×20 array of traps shown in the video micrograph. These are shown trapping 800 nm diameter polystyrene spheres dispersed in water. This figure was adapted with permission from ref. 31 © Elsevier Science Ltd. Bar, 10 μm .

and a polarization vector ϵ describing the field's orientation. A multi-beam interference hologram generally would modify both the amplitude and the phase of the input beam, with the amplitude modifications diverting power away from the optical traps. Fortunately, a variety of iterative optimization algorithms have been developed^{29–31} to create equivalent holographic beamsplitters that modify only the phase of the input beam. Such a phase-only diffractive optical element (DOE), also known as a kinoform, was used to create the 20×20 optical traps shown in Fig. 2.

Holographic optical tweezers really hit their stride when a computer-addressed spatial light modulator (SLM) was used to project sequences of trap-forming kinoforms in real time^{30–32}. An SLM imposes a prescribed amount of phase shift at each pixel in an array by varying the local optical path length. Typically, this is accomplished by controlling the local orientation of molecules in a layer of liquid crystal, although arrays of microelectromechanical (MEMS) mirrors are also becoming available for SLM applications. Slightly displacing the traps from one pattern to the next transfers particles along arbitrary three-dimensional (3D) trajectories^{30–32}, animating matter with light in much the same way that cartoons animate light with matter. Figure 3 shows this principle in action.

In a variation on this theme, the generalized phase contrast (GPC) technique converts a pattern of phase modulation across an SLM's face directly into the corresponding intensity modulation in the focal plane of the objective lens³³ and thus creates arbitrary planar trapping patterns. The conversion involves an annular phase plate similar to that used in phase contrast microscopy. This approach avoids the need to calculate holograms and thus is extremely efficient.

Figure 3 Polyesterene and silica spheres in two- and three-dimensional configurations of holographic optical tweezers created from a single laser beam with a computer-designed hologram of a single beam's wavefront. **a–c**, Thirty-six water-borne polystyrene spheres, 800 nm in diameter, are trapped in a plane and reconfigured with dynamic trapping patterns. Reproduced with permission from ref. 31 © Elsevier Science Ltd. Bar, 10 μm . **d,e**, Two 1 μm diameter silica spheres being moved past each other in two different planes. Reproduced with permission from ref. 30 © Elsevier Science Ltd. The different appearance of the spheres results from their different heights relative to the microscope's focal plane. **f–h**, Seven 1 μm diameter silica spheres being moved up and down through seven different planes. Reproduced with permission from ref. 31 © Elsevier Science Ltd.



The spatial resolution of existing SLMs currently limits the GPC technique to creating lateral traps rather than 3D optical tweezers, but GPC has still proved useful for rapidly organizing small objects in thin samples³⁴.

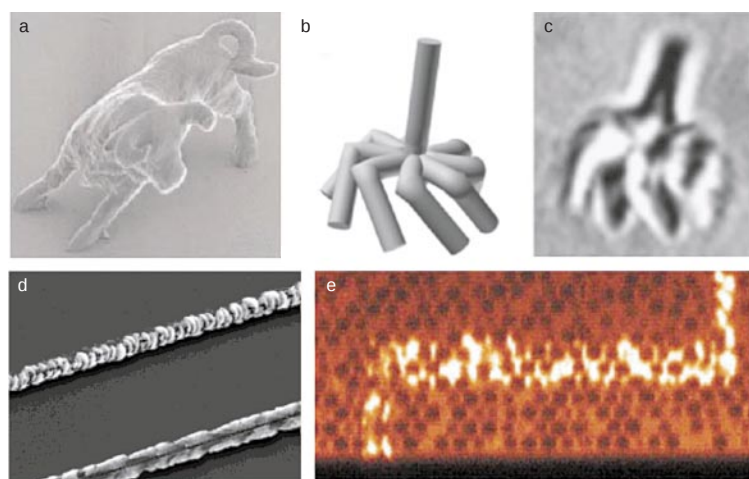
Even static arrays of optical traps have exciting and surprising applications. For example, an array of traps can continuously sort fluid-borne particles, acting much like a sieve. The array sorts particles on the basis of their different affinities for optical traps and for their affinity for the externally applied driving force. Inclining a regular array with respect to the driving force deflects the selected fraction so that it can be collected separately from the other, undeflected fraction³⁵. Unlike most sorting techniques that operate on discrete batches of sample, optical fractionation works continuously and can be dynamically optimized by adjusting the wavelength, intensity and geometry of the trap array. Moreover, because optical fractionation relies on the object's ability to hop from potential well to potential well, it is exponentially sensitive to particle size and so promises unparalleled size resolution³⁶.

An array of traps may also be viewed as a tailor-made potential energy landscape for interacting colloidal particles. Determining how strongly interacting systems evolve on modulated substrate

potentials is a classic problem in statistical physics, and colloids in modulated optical fields constitute a rare model system where microscopic interactions can be measured and controlled while macroscopic thermodynamic properties unfold^{37,38}. Insights obtained from studying optically modulated colloids are relevant to analogous systems such as atoms adsorbed on crystal surfaces, electrons passing through charge density waves and 2D electron gases, magnetic flux quanta passing through defects in type II superconductors and motor proteins translating along filaments in living cells. Early studies demonstrated that modulation along even one direction can freeze a 2D colloidal fluid^{39,40}. Deeper modulation actually melts the substrate-induced crystal⁴¹ by suppressing inter-row coupling⁴². More recent studies have demonstrated other intriguing behaviours, such as rotational melting in an array of multiply occupied traps^{38,43}, and have shed new light on the mechanisms by which magnetic flux lines invade superconductors^{37,38}.

Time-varying potential energy landscapes created with dynamic optical traps promise new insights into the operation of molecular motors by providing a powerful experimental system within which to study thermal ratchets²⁶ and related models in non-equilibrium statistical mechanics⁴⁴. Once perfected, such ratchet potentials will also

Figure 4 The diffraction-limited focus of an optical tweezer is ideal for spatially localized photochemistry. Examples include: **a**, the fabrication of a photopolymerized sculpture whose finest features are roughly 100 nm across (reproduced from ref. 84 © Macmillan Magazines Ltd); the fabrication of a light-driven turbine, shown as a solid model of a photopolymerized turbine with submicrometre features in **b** and as an actual turbine suspended in water in **c** (reproduced with permission from ref. 50 © American Institute of Physics); **d**, fine lines of MoS_2 deposited on a glass substrate by photoreduction of an aqueous salt solution (reproduced with permission from ref. 51 © Wiley-VCH); and **e**, a three-dimensional fluorescent polymer structure embedded in a colloidal crystal (reproduced from ref. 54 with permission from P. Braun).



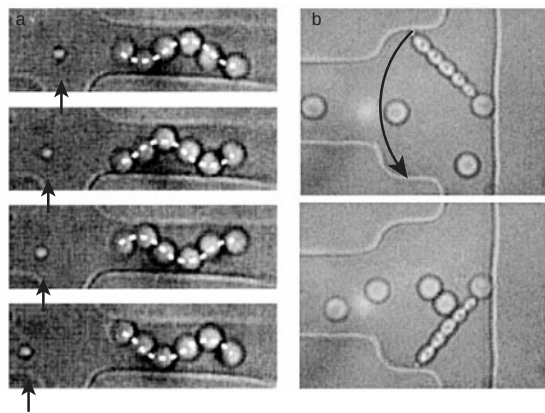


Figure 5 Optical pump and valve constructed of colloidal particles in microfluidic channels activated with optical tweezers. **a**, The flow of water created by the colloidal peristaltic pump is shown by the position of the tracer particle, which is indicated with an arrow (reproduced with permission from ref. 85 © American Association for the Advancement of Science). **b**, The colloidal valve flap is flipped with an optical tweezer and directs particles either downward (top) or upward (bottom). Reproduced with permission from ref. 55 © American Institute of Physics.

be useful for dynamically sorting mesoscopic objects and transporting them through tiny integrated laboratories for processing and testing⁴⁵.

All such studies provide valuable insights into how nature creates and exploits hierarchically organized structures. Once these principles are understood, they will be extraordinarily useful for creating new materials and devices to order. Until then, many of the most interesting three-dimensionally structured functional systems can be assembled using many optical tweezers operating in concert. Indeed, optical tweezers have an essentially unique ability to construct 3D heterostructures with features ranging in size from a few nanometres to a few millimetres. The real power of this approach only becomes apparent when optical trapping is combined with other techniques to create permanent structures with embedded functionality.

Nanofabrication with optical tweezers

Once assembled, tweezer-organized structures can be fixed in place, for instance, by sintering or gelling. The tweezers themselves can be used in this process. In particular, the intense illumination at an optical tweezer's focus is ideal for driving photochemical reactions in restricted volumes. If the reaction rate depends strongly on intensity, the resulting spatially resolved photochemistry can yield features smaller than the wavelength of light.

The first such application of optical tweezers involved the spatially resolved photo-oxidation of biological materials such as chromosomes^{46,47}. Essentially, a scalpel was created from light. Optical scalpels and scissors have been used for surgery on living cells^{15,48} as well as for ablating subwavelength structures into microscopic substrates⁴⁹.

Spatially resolved photochemistry using optical tweezers has been used to fabricate small complex 3D structures such as the examples shown in Fig. 4. Figure 4a–c shows 3D plastic structures created by multiphoton photopolymerization in scanned optical tweezers. The smallest features in Fig. 4a are about 100 nm across. The tiny turbine in Fig. 4b, c not only was created in this way but also was trapped and spun on its axis with an optical tweezer⁵⁰. Arrays of interlocking turbines and gears assembled and driven by light have already been demonstrated⁵⁰. Other photochemical transformations provide opportunities for optical tweezers to create 3D electronic and photonic structures. The fine lines of MoS₂ in Fig. 4d were patterned on glass by photoreduction of aqueous salts, and similar results have been obtained in silver, gold and oxidized copper⁵¹.

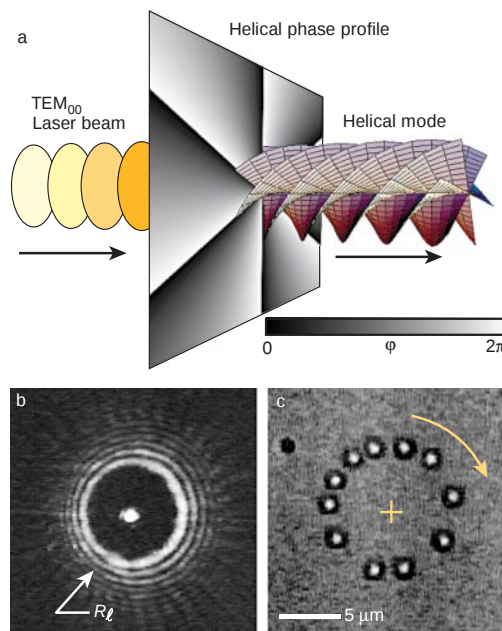


Figure 6 Optical vortices and optical spanners created from helical modes of light. **a**, The helical phase profile $\varphi(\rho) = \ell\theta$ converts a TEM₀₀ laser beam into a helical mode whose wavefronts resemble an ℓ -fold corkscrew. **b**, Rather than focusing to a point, a helical mode focuses to an optical vortex whose radius R_l is proportional to its pitch, ℓ . **c**, A single colloidal particle trapped in the optical vortex travels around its circumference, driven by the orbital angular momentum of the helical beam. This multiple exposure shows 11 stages, at 1/6 s intervals, in one 800 nm particle's transit. Reproduced with permission from ref. 67 © American Physical Society.

Beyond creating structures *de novo*, spatially resolved photochemistry can be used to modify pre-existing structures. The 3D optical waveguide structure in Fig. 4e demonstrates this principle. Here, a self-assembled crystal of colloidal silica spheres was perfused with a photosensitive precursor and selectively patterned with an optical tweezer to create the embedded polymer structure shown in Fig. 4e (ref. 52). Filling the gaps with a high-index material and then dissolving away the spheres and polymer would leave a tweezer-drawn waveguide pattern embedded in the otherwise self-assembled photonic crystal⁵³. This hybrid approach to creating hierarchically structured materials could sweep away many of the practical hurdles that have prevented self-assembled systems from making bigger inroads in photonics and electro-optical systems⁵⁴.

Using many optical tweezers to simultaneously organize prefabricated nanometre-scale parts and to stitch them together with spatially resolved photochemistry would yield a whole new category of hierarchically structured materials and devices. Such scale-spanning heterostructures would provide the building blocks for sensors, photonic devices and a host of other technologies. Hierarchically structured micromechanical systems hold similar promise for optomechanical and microfluidic applications. In this case, optical trapping also solves the outstanding problem of actuating such small devices. Some aspects of this solution involve the unusual and counterintuitive properties of traps created with newly discovered modes of light.

Optical actuators

Using conventional optical tweezers as actuators for micromachines is likely to speed up the adoption of lab-on-a-chip and related technologies for medical diagnostics, environmental testing and point-of-use microfabrication. Dynamic optical tweezers can both

organize and drive devices, such as the micrometre-scale hydraulic pump shown in Fig. 5a. The microscopic valve flap in Fig. 5b is an example of a photopolymerized colloidal heterostructure that is assembled and actuated with optical tweezers⁵⁵.

Modifying the wavefronts of the optical tweezers transforms them into whole new classes of optical traps, some of which have already found applications as actuators for unconventional micromachines. Some of the most useful of these are based on exotic modes of light whose properties have only recently been elucidated.

Figure 6a shows how the deceptively simple phase profile $\varphi(\rho) = \ell\theta$ transforms the parallel wavefronts of a TEM₀₀ laser mode into the corkscrew topology of a helical mode⁵⁶. Here, θ is the azimuthal angle around the optical axis and ℓ is an integer winding number (also known as the topological charge). The modified beam no longer focuses to a point, because the helical topology fosters destructive interference along the optical axis. Instead, it converges to a ring of light, as shown in Fig. 6b. The dark focus is suitable for trapping reflecting⁵⁷, absorbing⁵⁸ or low-dielectric-constant^{59,60} objects that would be damaged or repelled by conventional optical tweezers. Because such traps lack the radiation pressure from axial rays, they make more efficient traps for large dielectric objects than conventional optical tweezers do^{61–63}. Smaller dielectric particles are drawn to the ring's circumference, as shown in Fig. 6c.

What really distinguishes these ring-like optical traps is their ability to exert torques as well as forces^{57,58,64}. Just a decade ago, Allen demonstrated that each photon in a helical mode carries an orbital angular momentum $\ell\hbar$, where $\hbar$ is the quantum unit of angular momentum, in addition to its intrinsic spin angular momentum⁵⁶. This orbital angular momentum takes the form of a tangential component to the beam's linear momentum density that can be transferred to illuminated objects^{65–67}. A single colloidal microsphere is shown circulating around such a topological ring-trap under the influence of the optical angular momentum flux in the time-lapse photograph of Fig. 6c. Such toroidal torque-exerting traps have come to be known as optical vortices⁶⁸ or optical spanners⁶⁹, and they have potentially widespread technological applications. Studying the motion of objects in optical vortices has also provided valuable insights into the interplay of photon spin and orbital angular momentum^{57,64–67,70}, which have been useful in elucidating the quantum mechanical nature of helical beams.

An optical vortex's radius, R_ℓ , increases with topological charge^{67,71}; therefore, the intensity pattern can be tailored to different applications. For example, a properly scaled ring of light can be projected onto the teeth of a microfabricated gear, thereby creating a reliable micrometre-scale motor. The distributed drive made possible by projecting multiple optical vortices should also help to alleviate problems associated with friction in micromechanical systems. The spin angular momentum carried by circularly polarized optical tweezers similarly has been used to apply torques to birefringent components^{72–74}. Using an SLM to control the polarization of multiple optical tweezers opens up even more possibilities for developing extensive micromachines assembled and driven by light.

Some micromechanical applications may require no microfabrication at all. Rapidly circulating particles entrain flows that can mix and pump extremely small volumes of fluid. This solves a problem in microfluidic systems, whose laminar flows are ideal for transporting minuscule quantities of reagents but do not promote mixing when needed. Furthermore, the holographic optical tweezer technique can project multiple optical vortices, such as the 3×3 array in Fig. 7a, each with an individually specified intensity and topological charge³¹. Cooperative flows in such arrays can be reconfigured dynamically by modifying the trap-forming hologram, opening up the possibility of developing adaptive microfluidics on length scales ranging all the way down to tens of nanometres.

Other variations on this theme yield a family of distinct optical micromanipulators, each with its own applications. For example, superposing a helical beam on a conventional beam not only shows the helical wavefronts' structure, as in Fig. 7b, but also creates an

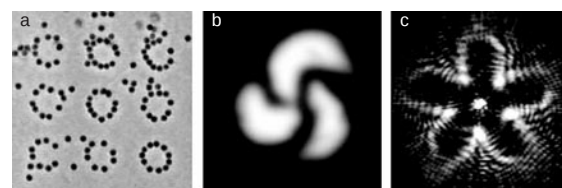


Figure 7 Generalizations of the optical vortex principle. **a**, The holographic optical tweezer technique can create arrays of optical vortices, each with individually specified intensity and topological charge. This 3×3 array of $\ell=30$ optical matrices is shown trapping 800 nm diameter polystyrene spheres. Reproduced with permission from ref. 31 © Elsevier Science Ltd. **b**, Trapped objects can be rotated within the interference pattern of an optical vortex and a plane wave. Such an optical rotator was created by interference of an $\ell=3$ helical mode with a plane wave. Reproduced with permission from ref. 75 © American Association for the Advancement of Science. **c**, An optical rotator created with a five-fold modulation of the helicity of an $\ell=60$ optical vortex. Reproduced with permission from ref. 67 © American Physical Society.

orientated intensity pattern that is useful for orientating asymmetric objects⁷⁵. Superposing instead a helical mode on its mirror-image counterpart creates 3D arrays of discrete traps that can be rotated arbitrarily in three dimensions by varying the beams' relative phase⁷⁶. Modulating the helical pitch of an optical vortex results in another class of optical rotators⁷⁷, an example of which appears in Fig. 7c. Further generalizations create intensity patterns related to the caustics seen at the bottom of swimming pools and which can move objects along complex trajectories transverse to the optical axis, all with static holograms and no moving parts. Other superpositions can focus on micrometre-scale dark regions surrounded by light on all sides known as optical bottles⁷⁸. These are useful for trapping very small dark-seeking objects, including clouds of ultracold atoms⁷⁸. Holographic arrays of optical bottles therefore should be useful for manipulating atoms⁷⁹, perhaps for quantum computing applications, and will help to extend pioneering efforts to apply optical tweezers to atomic physics^{80,81}.

Whereas azimuthal phase modulations extend optical tweezers into micromanipulators that are transverse to the optical axis, radial modulations create axial devices with an intriguing twist. The simplest non-trivial radial phase profile modification, $\varphi(\rho) = \gamma\rho$, transforms a TEM₀₀ beam into an approximation of a Bessel mode, a beam that propagates without diffracting even when focused to a wavelength-scale cross-section. The associated optical trap can extend for millimetres along the optical axis, as demonstrated in Fig. 8, and can precisely push particles over very large distances⁸³. The extended range of Bessel-beam arrays should increase the throughput of optical fractionation by orders of magnitude. Still more remarkably, Bessel beams are impervious to distortions by intervening particles and surfaces⁸⁴—they can reconstruct their wavefronts as they propagate away from disturbances. Combining the robustness of Bessel beams with the orbital angular momentum of helical modes yields optical devices that can reach deeply into complex systems to apply forces and torques where needed.

Future prospects

The ability to reach into the microscopic world dextrously and non-invasively at many points at once, to cut, assemble and transform with nanometre precision and submicrometre resolution and to do all these things with a single instrument promises revolutionary advances in many disciplines. The sections above highlight just a few of these advances. In particular, wavefront engineering provides a straightforward means to create large many optical traps in arbitrary 3D configurations, to move them freely and independently in three dimensions and to transform them into optical vortices, optical bottles, Bessel traps and a host of other all-optical tools.

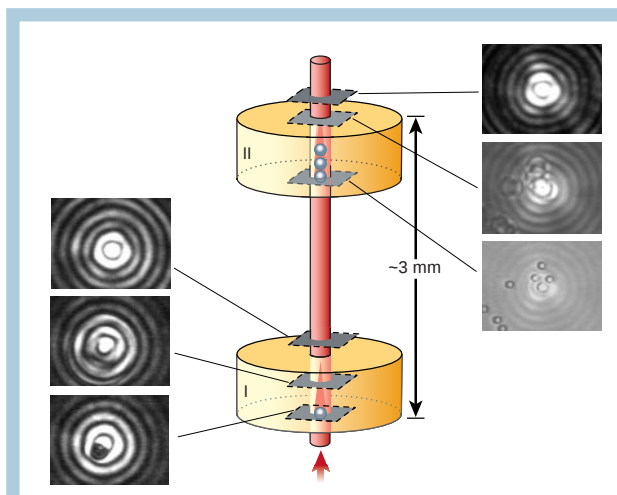


Figure 8 The radial phase profile $\varphi(\rho) = \gamma\rho$ creates a diffractionless Bessel beam that focuses to a long axial trap that can extend for millimetres. Here, the same beam is shown trapping multiple colloidal particles in two separate sample chambers separated by a distance of 3 mm. Bessel beams are impervious to distortions by intervening particles because wavefronts are reconstructed as they propagate away from disturbances. Left, a sphere is trapped between the central spot and the first ring of the Bessel beam in cell I. A little way above, the beam is distorted, but further above, the beam is no longer distorted. The process is repeated for the same beam in cell II in the planes on the right. Reproduced from ref. 83 © Macmillan Magazines Ltd.

As tools for biology, multifunctional optical traps will facilitate new approaches to cell sorting, macromolecular purification, intracellular surgery, embryonic testing and highly parallel drug screening as well as great many other possibilities. The same tools have immediate applications for organizing mesoscopic matter into heterogeneous, hierarchically structured 3D functional systems, such as photonic circuit elements, integrated sensor arrays and high-density data storage devices. Combining this organizational capability with optical-tweezer-based spatially resolved photochemistry suggests bright prospects for optically assembling new materials and devices with features ranging in size from nanometres to millimetres and beyond.

In micromechanics and microfluidics, appropriately sculpted wavefronts of light can easily control motions and flows on a length scale that has challenged other technologies. In so doing, optical micromachines should hasten the adoption of lab-on-a-chip devices for diagnostics, sensing, testing, pathology and drug discovery. The same wavefront-shaping techniques can sort and purify materials in these tiny flows and direct them towards further stages of purification and analysis. Optical testing and manufacturing, thus, could be highly integrated, with a single instrument providing flow, sorting, organization, synthesis and assembly. In all of these areas, the emerging generation of optical manipulation tools should help to bridge the chasm between our macroscopic world and applications based on the physics, chemistry and biology of microscopic systems. □

doi:10.1038/nature01935

- Ashkin, A., Dziedzic, J. M., Bjorkholm, J. E. & Chu, S. Observation of a single-beam gradient force optical trap for dielectric particles. *Opt. Lett.* **11**, 288–290 (1986).
- Ashkin, A. History of optical trapping and manipulation of small-neutral particle, atoms, and molecules. *IEEE J. Sel. Top. Quantum Elec.* **6**, 841–856 (2000).
- Ashkin, A. Forces of a single-beam gradient laser trap on a dielectric sphere in the ray optics regime. *Methods Cell Biol.* **55**, 1–27 (1998).
- Svoboda, K. & Block, S. M. Optical trapping of metallic Rayleigh particles. *Opt. Lett.* **19**, 930–932 (1994).
- Ke, P. C. & Gu, M. Characterization of trapping force on metallic Mie particles. *Appl. Opt.* **38**, 160–167 (1999).
- Ghislain, L. P., Switz, N. A. & Webb, W. W. Measurement of small forces using an optical trap. *Rev. Sci. Instr.* **65**, 2762–2768 (1994).
- Rohrbach, A. & Stelzer, E. H. K. Trapping forces, force constants, and potential depths for dielectric spheres in the presence of spherical aberrations. *Appl. Opt.* **41**, 2494–2507 (2002).

- Litvinov, R. I., Shuman, H., Bennett, J. S. & Weisel, J. W. Binding strength and activation state of single fibrinogen-integrin pairs on living cells. *Proc. Natl. Acad. Sci.* **99**, 7426–7431 (2002).
- Gittes, F. & Schmidt, C. F. Signals and noise in micromechanical measurements. *Methods Cell Biol.* **55**, 129–156 (1998).
- Gittes, F. & Schmidt, C. F. Interference model for back-focal-plane displacement detection in optical tweezers. *Opt. Lett.* **23**, 7–9 (1998).
- Pralle, A., Prummer, M., Florin, E. L., Stelzer, E. H. K. & Horber, J. K. H. Three-dimensional high-resolution particle tracking for optical tweezers by forward scattered light. *Microsc. Res. Tech.* **44**, 378–386 (1999).
- Svoboda, K., Mitra, P. P. & Block, S. M. Fluctuation analysis of motor protein movement and single enzyme-kinetics. *Proc. Natl. Acad. Sci.* **91**, 11782–11786 (1994).
- Bustamante, C., Smith, S. B., Liphardt, J. & Smith, D. Single-molecule studies of DNA mechanics. *Curr. Opin. Struct. Biol.* **10**, 279–285 (2000).
- Berns, M. W., Tadir, Y., Liang, H. & Tromberg, B. Laser scissors and tweezers. *Methods Cell Biol.* **55**, 71–98 (1998).
- Wright, G., Tucker, M. J., Morton, P. C., Sweitzer-Yoder, C. L. & Smith, S. E. Micromanipulation in assisted reproduction: A review of current technology. *Curr. Opin. Obstet. Gyn.* **10**, 221–226 (1998).
- Crocker, J. C. & Grier, D. G. Microscopic measurement of the pair interaction potential of charge-stabilized colloid. *Phys. Rev. Lett.* **73**, 352–355 (1994).
- Crocker, J. C. & Grier, D. G. When like charges attract: The effects of geometrical confinement on long-range colloidal interactions. *Phys. Rev. Lett.* **77**, 1897–1900 (1996).
- Ohshima, Y. N. Direct measurement of infinitesimal depletion force in a colloid-polymer mixture by laser radiation pressure. *Phys. Rev. Lett.* **78**, 3963–3966 (1997).
- Crocker, J. C., Matteo, J. A., Dinsmore, A. D. & Yodh, A. G. Entropic attraction and repulsion in binary colloids probed with a line optical tweezer. *Phys. Rev. Lett.* **82**, 4352–4355 (1999).
- Verma, R., Crocker, J. C., Lubensky, T. C. & Yodh, A. G. Attractions between hard colloidal spheres in semiflexible polymer solutions. *Macromolecules* **33**, 177–186 (2000).
- Yodh, A. G. *et al.* Entropically driven self-assembly and interaction in suspension. *Phil. Trans. R. Soc. A* **359**, 921–937 (2001).
- Wang, G. M., Sevick, E. M., Mittag, E., Searles, D. J. & Evans, D. J. Experimental demonstration of violations of the second law of thermodynamics for small systems and short time scales. *Phys. Rev. Lett.* **89**, 050601 (2002).
- Sasaki, K., Koshio, M., Misawa, H., Kitamura, N. & Masuhara, H. Pattern formation and flow control of fine particles by laser-scanning micromanipulation. *Opt. Lett.* **16**, 1463–1465 (1991).
- Sasaki, K., Fujiwara, H. & Masuhara, H. Optical manipulation of a lasing microparticle and its application to near-field microspectroscopy. *J. Vacu. Sci. Tech. B* **15**, 2786–2790 (1997).
- Mio, C., Gong, T., Terray, A. & Marr, D. W. M. Morphological control of mesoscale colloidal models. *Fluid Phase Equilibria* **185**, 157–163 (2001).
- Faucheux, L. P., Bourdieu, L. S., Kaplan, P. D. & Libchaber, A. J. Optical thermal ratchet. *Phys. Rev. Lett.* **74**, 1504–1507 (1995).
- Verma, R., Crocker, J. C., Lubensky, T. C. & Yodh, A. G. Entropic colloidal interactions in concentrated DNA solutions. *Phys. Rev. Lett.* **81**, 4004–4007 (1998).
- Dufresne, E. R. & Grier, D. G. Optical tweezer arrays and optical substrates created with diffractive optics. *Rev. Sci. Instr.* **69**, 1974–1977 (1998).
- Dufresne, E. R., Spalding, G. C., Dearing, M. T., Sheets, S. A. & Grier, D. G. Computer-generated holographic optical tweezer arrays. *Rev. Sci. Instr.* **72**, 1810–1816 (2001).
- Liesener, J., Reicherter, M., Haist, T. & Tiziani, H. J. Multi-functional optical tweezers using computer-generated holograms. *Opt. Commun.* **185**, 77–82 (2000).
- Curtis, J. E., Koss, B. A. & Grier, D. G. Dynamic holographic optical tweezers. *Opt. Commun.* **207**, 169–175 (2002).
- Reicherter, M., Haist, T., Wagemann, E. U. & Tiziani, H. J. Optical particle trapping with computer-generated holograms written on a liquid-crystal display. *Opt. Lett.* **24**, 608–610 (1999).
- Mogensen, P. C. & Glückstad, J. Dynamic array generation and pattern formation for optical tweezers. *Opt. Comm.* **175**, 75–81 (2000).
- Rodrigo, P. J., Eriksen, R. L., Daria, V. R. & Glückstad, J. Interactive light-driven and parallel manipulation of inhomogeneous particles. *Opt. Exp.* **10**, 1550–1556 (2002).
- Korda, P. T., Taylor, M. B. & Grier, D. G. Kinetically locked-in colloidal transport in an array of optical tweezers. *Phys. Rev. Lett.* **89**, 128301 (2002).
- Ladavac, K., Kasza, K. & Grier, D. G. Optical fractionation. *Phys. Rev. Lett.* (in the press).
- Korda, P. T., Spalding, G. C. & Grier, D. G. Evolution of a colloidal critical state in an optical pinning potential. *Phys. Rev. B* **66**, 024504 (2002).
- Mangold, K., Leiderer, P. & Bechinger, C. Phase transitions of colloidal monolayers in periodic pinning arrays. *Phys. Rev. Lett.* **90**, 158302 (2003).
- Chowdhury, A., Ackerson, B. J. & Clark, N. A. Laser-induced freezing. *Phys. Rev. Lett.* **55**, 833–836 (1985).
- Loudiyi, K. & Ackerson, B. J. Direct observation of laser-induced freezing. *Physica A* **184**, 1–25 (1992).
- Wei, X. *et al.* T cell activation studied by optical trap. *Biophys. J.* **74**, A378 (1998).
- Bechinger, C., Brunner, M. & Leiderer, P. Phase behavior of two-dimensional colloidal systems in the presence of periodic light fields. *Phys. Rev. Lett.* **86**, 930–933 (2001).
- Brunner, M. & Bechinger, C. Phase behavior of colloidal molecular crystals on triangular light lattices. *Phys. Rev. Lett.* **88**, 248302 (2002).
- Reimann, P. Brownian motors: Noisy transport far from equilibrium. *Phys. Rep.* **361**, 57–265 (2002).
- Koss, B. A. & Grier, D. G. Optical peristalsis. *Appl. Phys. Lett.* **82**, 3985–3987 (2003).
- Ashkin, A., Dziedzic, J. M. & Yamane, T. Optical trapping and manipulation of single cells using infrared-laser beams. *Nature* **330**, 608–609 (1987).
- Liang, H., Wright, W. H., Cheng, S., He, W. & Berns, M. W. Micromanipulation of chromosomes in PtK2 cells using laser microsurgery (optical scalpel) in combination with laser-induced optical force (optical tweezers). *Exp. Cell Res.* **204**, 110–120 (1992).
- Bayles, C. J., Aist, J. R. & Berns, M. W. The mechanics of anaphase-B in *A. basidiomycete* as revealed by laser microbeam microsurgery. *Exp. Mycol.* **17**, 191–199 (1993).
- Fuhr, G. *et al.* Processing of micro-particles by UV laser irradiation in a field cage. *Appl. Phys. A* **69**, 611–616 (1999).
- Galajda, P. & Ormos, P. Complex micromachines produced and driven by light. *Appl. Phys. Lett.* **78**, 249–251 (2001).

51. Lachish-Zalait, A., Zbaida, D., Klein, E. & Elbaum, M. Direct surface patterning from solutions: Localized microchemistry using a focused laser. *Adv. Funct. Mat.* **11**, 218–223 (2001).
52. Lee, W. M., Pruzinsky, S. A. & Braun, P. V. Multi-photon polymerization of waveguide structures within three-dimensional photonic crystals. *Adv. Mat.* **14**, 271–274 (2002).
53. Ioannopoulos, J. D., Meade, R. D. & Winn, J. N. *Photonic Crystals* (Princeton Univ., Princeton, 1995).
54. Taton, T. A. & Norris, D. J. Device physics: Defective promise in photonics. *Nature* **416**, 685–686 (2002).
55. Terray, A., Oakey, J. & Marr, D. W. M. Fabrication of linear colloidal structures for microfluidic applications. *Appl. Phys. Lett.* **81**, 1555–1557 (2002).
56. Allen, L., Beijersbergen, M. W., Spreeuw, R. J. C. & Woerdman, J. P. Orbital angular-momentum of light and the transformation of Laguerre-Gaussian laser modes. *Phys. Rev. A* **45**, 8185–8189 (1992).
57. O’Neil, A. T. & Padgett, M. J. Three-dimensional optical confinement of micron-sized metal particles and the decoupling of the spin and orbital angular momentum within an optical spanner. *Opt. Commun.* **185**, 139–143 (2000).
58. Rubinsztein-Dunlop, H., Nieminen, T. A., Friese, M. E. J. & Heckenberg, N. R. Optical trapping of absorbing particles. *Adv. Quantum Chem.* **30**, 469–492 (1998).
59. Gahagan, K. T. & Swartzlander, G. A. Trapping of low-index microparticles in an optical vortex. *J. Opt. Soc. Am. B* **15**, 524–534 (1998).
60. Gahagan, K. T. & Swartzlander, G. A. Simultaneous trapping of low-index and high-index microparticles observed with an optical-vortex trap. *J. Opt. Soc. Am. B* **16**, 533–537 (1999).
61. Ashkin, A. Forces of a single-beam gradient laser trap on a dielectric sphere in the ray optics regime. *Biophys. J.* **61**, 569–582 (1992).
62. Simpson, N. B., McGloin, D., Dholakia, K., Allen, L. & Padgett, M. J. Optical tweezers with increased axial trapping efficiency. *J. Mod. Opt.* **45**, 1943–1949 (1998).
63. O’Neil, A. T. & Padgett, M. J. Axial and lateral trapping efficiency of Laguerre-Gaussian modes in inverted optical tweezers. *Opt. Commun.* **193**, 45–50 (2001).
64. He, H., Friese, M. E. J., Heckenberg, N. R. & Rubinsztein-Dunlop, H. Direct observation of transfer of angular momentum to absorptive particles from a laser beam with a phase singularity. *Phys. Rev. Lett.* **75**, 826–829 (1995).
65. Allen, L., Padgett, M. J. & Babiker, M. The orbital angular momentum of light. *Prog. Opt.* **39**, 291–372 (1999).
66. O’Neil, A. T., MacVicar, I., Allen, L. & Padgett, M. J. Intrinsic and extrinsic nature of the orbital angular momentum of a light beam. *Phys. Rev. Lett.* **88**, 053601 (2002).
67. Curtis, J. E. & Grier, D. G. Structure of optical vortices. *Phys. Rev. Lett.* **90**, 133901 (2003).
68. Gahagan, K. T. & Swartzlander, G. A. Optical vortex trapping of particles. *Opt. Lett.* **21**, 827–829 (1996).
69. Simpson, N. B., Allen, L. & Padgett, M. J. Optical tweezers and optical spanners with Laguerre-Gaussian modes. *J. Mod. Opt.* **43**, 2485–2491 (1996).
70. Simpson, N. B., Dholakia, K., Allen, L. & Padgett, M. J. Mechanical equivalence of spin and orbital angular momentum of light: An optical spanner. *Opt. Lett.* **22**, 52–54 (1997).
71. Padgett, M. J. & Allen, L. The Poynting vector in Laguerre-Gaussian modes. *Opt. Commun.* **121**, 36–40 (1995).
72. Higurashi, E., Sawada, R. & Ito, T. Optically induced angular alignment of trapped birefringent micro-objects by linearly polarized light. *Phys. Rev. E* **59**, 3676–3681 (1999).
73. Higurashi, E., Sawada, R. & Ito, T. Optically driven angular alignment of microcomponents made of in-plane birefringent polyimide film based on optical angular momentum transfer. *J. Microtech. Microeng.* **11**, 140–145 (2001).
74. Friese, M. E. J., Rubinsztein-Dunlop, H., Gold, J., Hagberg, P. & Hanstorp, D. Optically driven micromachine elements. *Appl. Phys. Lett.* **78**, 547–549 (2001).
75. Paterson, L. *et al.* Controlled rotation of optically trapped microscopic particles. *Science* **292**, 912–914 (2001).
76. MacDonald, M. P. *et al.* Creation and manipulation of three-dimensional optically trapped structures. *Science* **296**, 1101–1103 (2002).
77. Curtis, J. E. & Grier, D. G. Modulated optical vortices. *Opt. Lett.* **28**, 872–874 (2003).
78. Arlt, J. & Padgett, M. J. Generation of a beam with a dark focus surrounded by regions of higher intensity: The optical bottle beam. *Opt. Lett.* **25**, 191–193 (2000).
79. McGloin, D., Spalding, G. C., Melville, H., Sibbet, W. & Dholakia, K. Applications of spatial light modulators in atom optics. *Opt. Exp.* **11**, 158–166 (2003).
80. Gustavson, T. L. *et al.* Transport of Bose-Einstein condensates with optical tweezers. *Phys. Rev. Lett.* **88**, 020401 (2002).
81. Chikkatur, A. P. *et al.* A continuous source of Bose-Einstein condensed atoms. *Science* **296**, 2193–2195 (2002).
82. Arlt, J., Garces-Chavez, V., Sibbet, W. & Dholakia, K. Optical micromanipulation using a Bessel light beam. *Opt. Commun.* **197**, 239–245 (2001).
83. Garces-Chavez, V., McGloin, D., Melville, H., Sibbet, W. & Dholakia, K. Simultaneous micromanipulation in multiple planes using a self-reconstructing light beam. *Nature* **419**, 145–147 (2002).
84. Kawata, S., Sun, H.-B., Tanaka, T. & Takada, K. Finer features for functional microdevices. *Nature* **412**, 697–698 (2001).
85. Terray, A., Oakey, J. & Marr, D. W. M. Microfluidic control using colloidal devices. *Science* **296**, 1841–1844 (2002).

Discretizing light behaviour in linear and nonlinear waveguide lattices

Demetrios N. Christodoulides¹, Falk Lederer² & Yaron Silberberg³

¹School of Optics/CREOL, University of Central Florida, Orlando, Florida 32816-2700, USA

²Friedrich-Schiller-Universität Jena, Max-Wien-Platz 1, 07743 Jena, Germany

³Department of Physics of Complex Systems, The Weizmann Institute of Science, 76100 Rehovot, Israel

Light propagating in linear and nonlinear waveguide lattices exhibits behaviour characteristic of that encountered in discrete systems. The diffraction properties of these systems can be engineered, which opens up new possibilities for controlling the flow of light that would have been otherwise impossible in the bulk: these effects can be exploited to achieve diffraction-free propagation and minimize the power requirements for nonlinear processes. In two-dimensional networks of waveguides, self-localized states—or discrete solitons—can travel along ‘wire-like’ paths and can be routed to any destination port. Such possibilities may be useful for photonic switching architectures.

The ability to mould the flow of light in photonic periodic structures is a fundamental issue of scientific and practical importance. Optical waves propagating in such an environment experience a spatially periodic refractive index distribution and therefore behave in a way that is analogous to electrons travelling through a semiconductor crystal. A vital aspect associated with wave motion in photonic lattices is that, on many occasions, light can exhibit behaviour that is characteristic of that encountered in discrete systems. Such a case arises when the Floquet–Bloch modes of an optical field can be approximated by a finite, discrete set of bound states or modes. As a result, the underlying field evolution effectively becomes ‘discretized’. Arrays or lattices of evanescently coupled waveguides are prime examples of structures in which discrete optical dynamics can be observed and investigated. These arrays consist of equally spaced identical waveguide elements or sites (Fig. 1a, b) and, therefore, possess all the essential characteristics of a photonic crystal structure (Brillouin zones, allowed and forbidden bands and so on). In this physical setting, light ‘hops’ from site to site through optical tunnelling or coupling, and, in doing so, it profoundly alters its diffraction characteristics. The discrete diffraction properties of one-dimensional (1D) waveguide chains were first theoretically addressed in 1965¹ and experimentally observed in gallium arsenide arrays a few years later². At the time, however, it was by no means clear how one could either take advantage of or suppress this peculiar discrete diffraction process, and, as a result, the field remained dormant for several years.

In 1988, the idea that light could trap itself in nonlinear optical waveguide arrays or lattices was suggested³. This self-localization can take place provided that the on-site nonlinearity exactly balances the discrete diffraction arising from linear coupling effects among adjacent waveguides. As a consequence, the optical field is confined within only a few waveguide sites, and it forms what is better known as a discrete soliton (a nonlinear particle-like wavepacket, the shape of which remains invariant during propagation). The same discrete evolution equation that governs optical nonlinear waveguide lattices is also believed to describe quantum excitations in α -helical proteins—an idea first suggested in the early 70s^{4,5}. This link between optics and bioenergetics hints

at the exciting possibility that biological processes can be mimicked entirely by light.

The prospect of observing discrete solitons (DSs) in nonlinear optical waveguide arrays sparked a flurry of important theoretical studies in the 90s; all of them revealed fundamental differences between these discrete systems and their continuum counterparts^{6–9}. This effort culminated in the first successful experimental observation¹⁰ of optical DSs in AlGaAs waveguide arrays (Fig. 1a). Recent advances in microfabrication technology that allowed the fabrication of precision-made defect-free waveguide lattices made this possible. These initial experiments were followed by a series of breakthroughs, including the first observation of anomalous diffraction and diffraction management^{11,12}, Bloch oscillations and Floquet–Bloch solitons^{13–15}, DSs in nonlinear quadratic¹⁶ arrays and photorefractive^{17,18} arrays, to mention a few.

In this article, we review the current state of this emerging and exciting field. We emphasize that waveguide arrays can serve as archetypal structures in which the potential of linear and nonlinear discrete interactions can be assessed for future optical technologies. These types of interaction are universal: they are expected to occur in a variety of periodic structures such as photonic crystals^{19–21} and photonic crystal fibres²². DSs, the natural modes of any nonlinear lattice, can also be thoroughly investigated in waveguide arrays as a prelude to analysis in more complex nonlinear optical crystal systems. Furthermore, progress in this area is expected to shed light on other areas of science^{23–26}, where similar discrete-like dynamics are encountered, such as, Bose–Einstein condensates trapped in optical periodic potentials^{25,26}.

Linear diffraction properties of waveguide arrays

Optics deals with electromagnetic waves that are continuous functions of space and time. There are situations, however, where the evolution of an optical field can be represented as a discrete problem. As previously mentioned, one such simple and important case is that of light in a coupled 1D waveguide array^{1–3}. In a waveguide array, a large number (infinite in principle) of single-mode channel waveguides are laid next to each other in such a way that their individual modes overlap. As a result, every waveguide is coupled to its nearest neighbours, and, consequently, light propagation in such waveguide arrays is drastically affected. In this case, the optical mode electric field amplitude, E_n , at array site n

evolves according to the following linear discrete differential equation: $i(dE_n/dz) + \kappa(E_{n-1} + E_{n+1}) = 0$. In this equation, the first term describes the modal field evolution along the z direction (parallel to the waveguides), and the second one accounts for the evanescent field coupling, κ , of site n to its neighbouring waveguides.

To further understand the peculiarities of light propagation in linear waveguide arrays, let us first consider the simple case where only one waveguide is excited at the input. Under these conditions, the light intensity redistributes itself within the array as shown in Fig. 1c. Note that most of the light is concentrated in two distinct outer lobes. This is in contrast to the diffraction process that occurs in a bulk system, where most of the power is concentrated around the centre. This indicates that such discrete diffraction is profoundly different and has no analogue in continuous systems.

Discrete diffraction, however, has much more to offer. Unlike a bulk system, where the magnitude of diffraction is fixed at a given wavelength, in arrays, discrete diffraction can be engineered. To appreciate this, consider the dispersion relation of the governing discrete evolution equation, which is given by $k_z = 2\kappa \cos(k_x D)$ (ref. 3), where k_z is the propagation eigen value (or longitudinal $\mathbf{k}$ -vector) and $k_x D$ represents the so-called Bloch momentum (where D is the spacing between waveguide sites)³. The Bloch momentum vector is determined by the angle at which the beam is launched into the array^{11,12}. It is important to emphasize that in coupled arrays the dispersion curve is sinusoidal and not parabolic as it is in continuous media. Because the diffraction process is proportional to the curvature (or the second derivative) of the dispersion relation, one can conclude that diffraction is normal for $0 < k_x D < \pi/2$ (normal diffraction), it reverses its sign for Bloch momenta $\pi/2 < k_x D < \pi$ (anomalous diffraction) and it becomes zero at $k_x D = \pi/2$ (refs 11,12) as further explained in Fig. 2. These peculiarities led to a renewed interest in the linear properties of discrete optical structures. In particular, diffraction can be managed or engineered by concatenating sections with different signs of diffraction¹¹. Finally, the concept of discrete diffraction can be extended to higher-order Floquet–Bloch bands, as will be described in a later section of this review.

Photonic Bloch oscillations

As a direct consequence of the discrete translational symmetry, the underlying physical mechanism leading to discrete diffraction and refraction in homogeneous arrays is the band structure of the dispersion relation. Thus, the question naturally arises of how the field dynamics change if the waveguide lattice becomes inhomogeneous, for example, through introduction of a linear variation in the effective refractive index across the array. This brings forth a fundamental issue in discrete systems first addressed by Bloch and Zener 70 years ago^{27,28}. They raised the question of how do electrons behave in a crystal lattice when a DC field (linear potential) is applied, and they predicted the occurrence of periodic oscillations, later termed Bloch oscillations. These oscillations are caused by field-induced acceleration of electrons travelling in this tilted periodic potential. When the particle Bloch momentum reaches the Brillouin zone boundary, it is Bragg reflected and subsequently decelerated by the DC field until it comes to rest, completing one period of oscillation. Thus Bloch oscillations are the result of discreteness and lead to linear localization of excitation rather than unlimited spreading. The existence of these oscillations remained controversial for a long time and was finally confirmed when Bloch oscillations of electrons in semiconductor superlattices were experimentally observed²⁹. Inspired by these results, similar phenomena were searched for in other disciplines that involve periodic systems, for example, cold atoms and Bose–Einstein condensates in optical lattices and photons in periodic optical systems such as waveguide arrays.

An important advantage optical systems offer is the ability to directly visualize effects such as Bloch oscillations. In waveguide arrays, optical Bloch oscillations are possible so long as the refractive

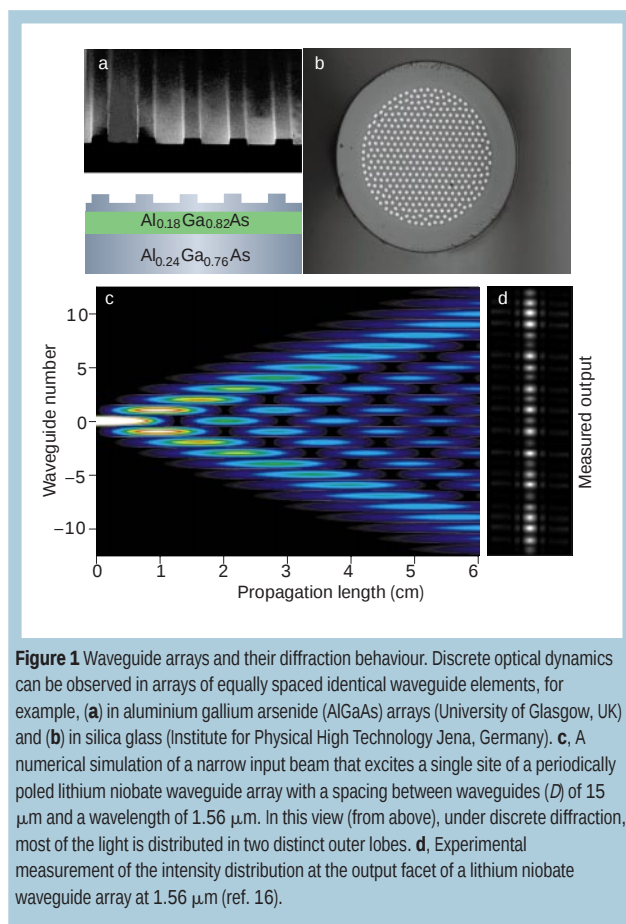


Figure 1 Waveguide arrays and their diffraction behaviour. Discrete optical dynamics can be observed in arrays of equally spaced identical waveguide elements, for example, (a) in aluminium gallium arsenide (AlGaAs) arrays (University of Glasgow, UK) and (b) in silica glass (Institute for Physical High Technology Jena, Germany). c, A numerical simulation of a narrow input beam that excites a single site of a periodically poled lithium niobate waveguide array with a spacing between waveguides (D) of 15 μm and a wavelength of 1.56 μm . In this view (from above), under discrete diffraction, most of the light is distributed in two distinct outer lobes. d, Experimental measurement of the intensity distribution at the output facet of a lithium niobate waveguide array at 1.56 μm (ref. 16).

index distribution varies linearly with site n ³⁰. In this case, the discrete field amplitudes E_n obey $i(dE_n/dz) + \alpha n E_n + \kappa(E_{n-1} + E_{n+1}) = 0$, where the array discrete evolution equation has been modified to account for the linear index tilt αn , with α being the strength of the 'linear potential'. This model exhibits localized solutions of identical shape. These are the so-called Wannier–Stark states, which have equally spaced eigen values (the Wannier–Stark ladder). Thus, any input distribution excites a certain set of Wannier–Stark states, and the superposition of these confined states leads to a periodic recurrence of the input field after a propagation distance $z = 2\pi m/\alpha$, where m is an integer. This wavepacket recurrence phenomenon was experimentally verified in an AlGaAs array consisting of waveguides of varying width¹³ and in a thermo-optic polymer array of identical waveguides¹⁴ as explained in Fig. 3.

Discrete solitons

Optical DSs are, in general, self-trapped entities of a finite extent that exist in coupled periodic nonlinear lattices³. They are self-localized wavepackets whose energy resides primarily in distinct waveguide array sites (hence discrete) and exist through a balance of discrete diffraction/coupling effects and material nonlinearity. Unlike their continuous counterparts, DSs represent collective excitations of the nonlinear chain (array), and, as a result, they exhibit a host of interesting characteristics that are absent in the bulk³¹.

Depending on the nature of the underlying nonlinear process, different families of discrete solitons are possible. To date, arrays made from materials with intensity-dependent Kerr¹⁰, quadratic¹⁶ or photorefractive nonlinearities¹⁸ have been shown to support such discrete localized states. For a given material, the nonlinear mechanism can be either self-focusing or defocusing⁶. In the self-

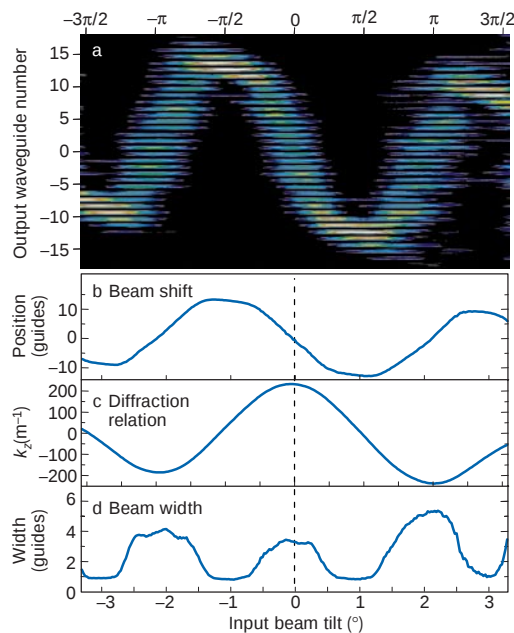


Figure 2 Phase difference between adjacent waveguides ($k_x D$). **a**, Displacement of a Gaussian beam launched in a 6 cm long polymer array at 0.633 μm (ref. 12). The beam displacement as measured experimentally at the output facet of the array is shown as a function of the initial beam tilt or phase difference between adjacent waveguides ($k_x D$). This displacement or shift is proportional to the transverse ‘velocity’ with which a beam travels across the array lattice. Because the group velocity happens to be the first derivative of the sinusoidal dispersion relation $k_x = 2\kappa \cos(k_x D)$ (see **b**), the beam exhibits zero displacement at the base of the band ($k_x D = 0$) and at the edge of the Brillouin zone ($k_x D = \pm\pi$). By contrast, the beam shift is at its maximum at $k_x D = \pm\pi/2$ (middle of Brillouin zone), as illustrated in **b**. From these data, the sinusoidal dispersion relation (see **c**) for k_x can be determined through integration. The diffraction experienced by a beam in the array is proportional to the second derivative of the dispersion relation $d^2 k_x / dk_x^2$ and is therefore normal around the base of the band ($k_x D = 0$) and anomalous close to the edge of the Brillouin zone ($k_x D = \pm\pi$) (ref. 58). For a Bloch momentum $k_x D = \pi/2$, diffraction is completely arrested, thus a wavepacket can propagate in the linear array with virtually no diffraction. **d**, The measured beamwidth (proportional to the array diffraction) at the output of the array as a function of the input beam tilt or phase difference $k_x D$ is shown. It is easily seen that beam spreading is minimal at $k_x D = \pm\pi/2$ and attains a maximum at $k_x D = 0, \pm\pi$.

focusing case, the refractive index increases locally with intensity, whereas in the defocusing cases it decreases.

Before we discuss the various DS families, it may be beneficial to understand the physics behind DS formation. To comprehend this nonlinear self-trapping effect, it is important to reiterate some of the most significant aspects of linear wave propagation in array lattices. In the linear regime, the optical field travelling in a waveguide array is subjected to a periodic potential, and, as a result, the dispersion relation (between the propagation eigen value and the transverse $\mathbf{k}$ -vector) is organized as a succession of allowed bands and bandgaps, as shown in Fig. 4a, within the Brillouin zone¹⁵. Waves are only able to ‘travel’ if their eigen values fall within an allowed band (Floquet–Bloch modes), whereas wavefunctions with propagation constants lying in the bandgaps decay exponentially in the transverse direction.

Under nonlinear conditions (such as in Kerr-like media), however, the optical field perturbs, through nonlinearity, the refractive index, thus, inducing a defect within an otherwise perfect waveguide lattice. Consequently, the eigen value of this nonlinear defect-like state moves into a bandgap, and, therefore, its wave function decays

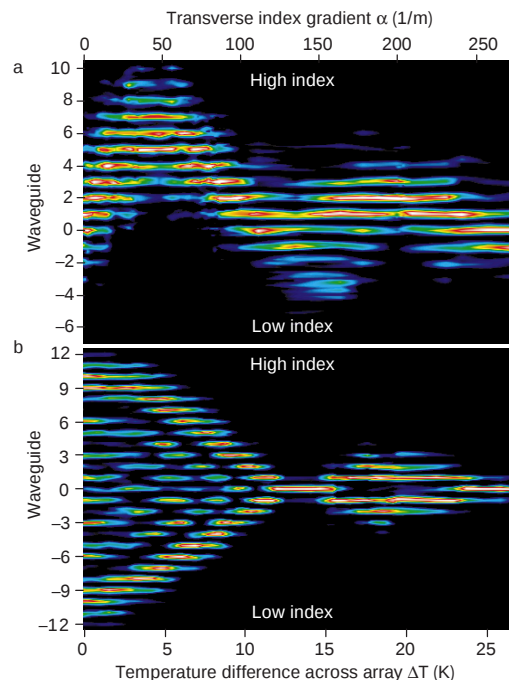


Figure 3 Experimental results depicting optical Bloch oscillations in a 4.5 cm long thermo-optically tuned polymer array at a wavelength of 0.633 μm (ref. 14). **a**, The intensity distribution at the output facet of the array versus the index gradient α or the temperature difference ΔT when the array is initially excited with a Gaussian beam. By increasing the temperature difference ΔT the oscillating transverse motion of the output beam can be measured. The decrease in the maximum elongation between $\alpha = 70$ to 210 m^{-1} is due to the stronger localization of Wannier–Stark states. At $\alpha = 140, 280 \dots \text{m}^{-1}$, full Bloch oscillations occur and the initial Gaussian distribution is recovered. **b**, Similar results were observed when only one waveguide was initially excited. When $\Delta T = 0$, the familiar discrete diffraction pattern (Fig. 1c, d) is recovered, whereas for integer multiples of the Bloch period $\alpha = 140, 280, \dots \text{m}^{-1}$, the beam narrows. Similar effects were observed in AlGaAs arrays at low intensities by monitoring the output intensity for different sample lengths¹³.

rapidly away from the perturbed region. Hence, the field becomes self-localized and a DS forms. Note that as the power of this optical field increases, the deeper its propagation eigen value moves into the gap, resulting in a more confined and transversely immobile DS. If the waveguide sites are sufficiently separated, then the Floquet–Bloch functions belonging to the first allowed band of the array could be described through coupled-mode theory or tight-binding approximation³². In this formalism, the local bound states or waveguide modes discretely interact with each other through evanescent coupling. In this case, for Kerr nonlinear arrays, the electric field amplitudes, E_n , of the modal fields involved obey the so-called discrete nonlinear Schrödinger equation³: $i(dE_n/dz) + \kappa(E_{n+1} + E_{n-1}) + |E_n|^2 E_n = 0$. This is a generalization of the discrete evolution equation, and the last term describes the on-site nonlinearity. Like other discrete nonlinear lattice equations, such as the Fermi–Pasta–Ulam problem, the Toda equation or the Ablowitz–Ladik equation³³, this equation has DS solutions (see ref. 6 for more details). In general, DS solutions with zero transverse velocity can be subdivided into two basic categories: (i) those with zero Bloch momentum $k_x D = 0$ (at the base of the band) and (ii) those at the edge of the band within the Brillouin zone where the Bloch momentum is $k_x D = \pi$. For example, at the base of the first band of Fig. 4a, the curvature of the dispersion curve is such that discrete diffraction is normal. As a result, in-phase bright DSs exist in

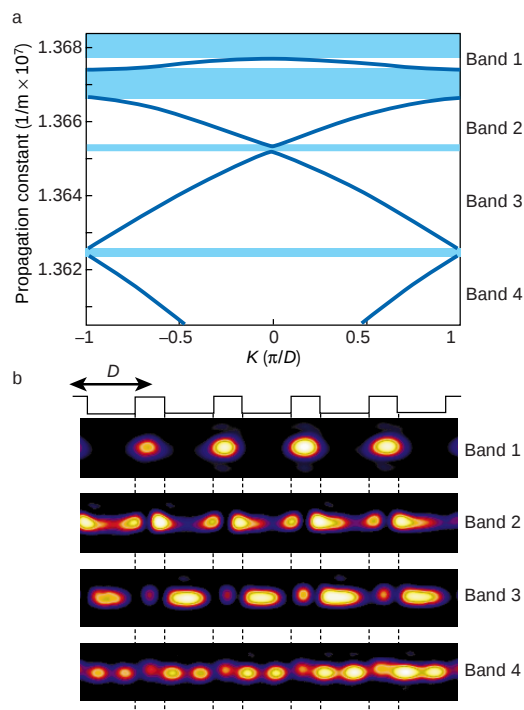


Figure 4 Array band structure and the associated Floquet–Bloch modes. **a**, Reduced bandgap diagram of a typical waveguide array (folded into the first Brillouin zone), where the Floquet–Bloch mode propagation constant or eigen value is plotted as a function of the Bloch momentum wavevector (K). The shaded regions represent the gaps, and the blue curves between the gaps. The first four allowed bands are shown. **b**, Highly magnified images of the excited Floquet–Bloch modes corresponding to the first four bands obtained experimentally in ref. 15 at a wavelength of $1.55 \mu\text{m}$ and using an AlGaAs array with $D=11 \mu\text{m}$.

arrays with self-focusing nonlinearities. By contrast, in defocusing systems, dark in-phase DSs exist instead. At the edge of the band where the discrete diffraction is anomalous (negative curvature), the situation is reversed. In this case, dark solitons exist in self-focusing arrays and bright solitons in defocusing ones^{9,34,35}. This time, however, the modal fields happen to be π out of phase or staggered.

Figure 5a shows the first experimental observation of an in-phase bright discrete soliton in a self-focusing AlGaAs waveguide array¹⁰. At low optical powers, the optical field discretely diffracts in the array, as previously discussed. At higher powers, however, the optical field self-localizes, which provides a clear indication of discrete soliton formation. In addition, dark staggered solitons ($k_x D = \pi$) have been recently observed in the same waveguide array by launching the beam at an appropriate angle into the system³⁵. Similarly, transport dynamics of DSs, as well as effects arising from the so-called Pöschl–Teller potential, have also been experimentally investigated³⁶. Very recently, Floquet–Bloch solitons have been successfully observed in waveguide arrays¹⁵. These solitons ‘reside’ in the higher-band gaps of Fig. 4a and therefore involve Floquet–Bloch modes from higher bands. Figure 4b shows the Floquet–Bloch structure of these higher-order soliton modes. Within the past few months, DSs have also been observed in biased photorefractive crystals¹⁷. In this system, both in-phase bright and staggered bright DSs have been observed by using the versatility of photorefractive nonlinearities. Other interesting DS families, such as vector³⁷, diffraction-managed³⁸ and gap solitons in diatomic lattices³⁹, are predicted to exist in waveguide arrays. Furthermore,

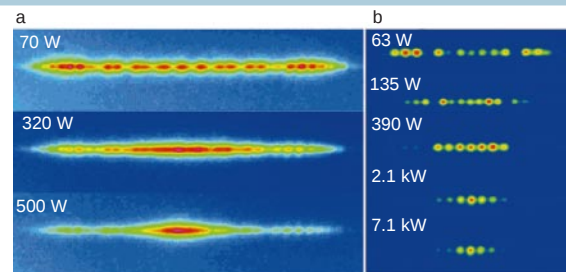


Figure 5 Experimental observation of discrete solitons. **a**, Intensity images at the output of a Kerr AlGaAs array with a waveguide spacing of $D=8 \mu\text{m}$ and operated at $1.53 \mu\text{m}$ observed at various power levels. At low powers (70 W), the beam discretely diffracts, whereas at 500 W, a discrete soliton forms¹⁰. **b**, Intensity images from a periodically poled lithium niobate (PPLN) array with $D=15 \mu\text{m}$ and used at $1.56 \mu\text{m}$ (ref. 16). At a power level of 7.1 kW, a discrete soliton forms.

interaction forces among discrete solitons⁴⁰ have been experimentally observed in AlGaAs arrays under various input conditions⁴¹; discrete vector solitons were also observed⁴².

During the past several years, a great deal of attention has been paid to the behaviour of spatial solitons in quadratic nonlinear media³¹. Unlike their Kerr counterparts, quadratic spatial solitons are symbiotically or dichromatically localized objects, because their fundamental and harmonic waves are mutually locked and do not exchange energy. Moreover, in these systems the wave vector mismatch between both constituents introduces an additional degree of freedom with regard to soliton formation. Discrete quadratic solitons were predicted⁴³ in the late 90s, and since then they have been the focus of ongoing research⁶. Yet, it is only recently that the technology of Ti-indiffused (doped) periodically poled lithium niobate (PPLN) waveguides has matured to the point that quadratic waveguide arrays can be fabricated. The first experimental results confirming the existence of discrete quadratic solitons¹⁶ are shown in Fig. 5b.

Discrete solitons in higher dimensions and other settings

So far we have focused on DSs in a 1D environment. Nevertheless, these entities can exist in higher dimensions (2D and 3D) so long as a nonlinear lattice system is present to host them. Indeed, there are many exciting DS features that can only occur in higher dimensions. These include vortex lattice solitons⁴⁴, bright discrete solitons carrying angular momenta and slow light DSs in 3D photonic crystal networks^{45,46}, just to mention a few. Even though creating such lattices in multiple dimensions has always been technologically challenging, recent developments are now promising to alleviate the fabrication difficulties that have so far inhibited progress in this field.

DSs were predicted⁴⁷ and subsequently experimentally observed¹⁸ for the first time in biased photorefractive crystals such as Strontium–Barium–Niobate:75 (SBN). A key element that led to the success of this particular experiment was the large electro-optic anisotropy of SBN:75, which allowed virtually linear propagation in one polarization and highly nonlinear propagation in the other. By exploiting this, a 2D waveguide lattice can be optically induced by interfering plane wave pairs polarized along the linear axis. By contrast, a beam polarized along the other axis will experience strong nonlinearity as well as the optically induced periodic potential, thus forming a 2D discrete optical soliton. Moreover, owing to the high nonlinearity of photorefractive crystals, new families of 2D DSs can be observed at milliwatt power levels. This technique offers considerable flexibility in that the same photorefractive waveguide lattice can be either self-focusing or defocusing depending on the crystal bias polarity. Last but not least, a rich variety of optically induced 2D lattices is possible in these systems, such as rectangular,

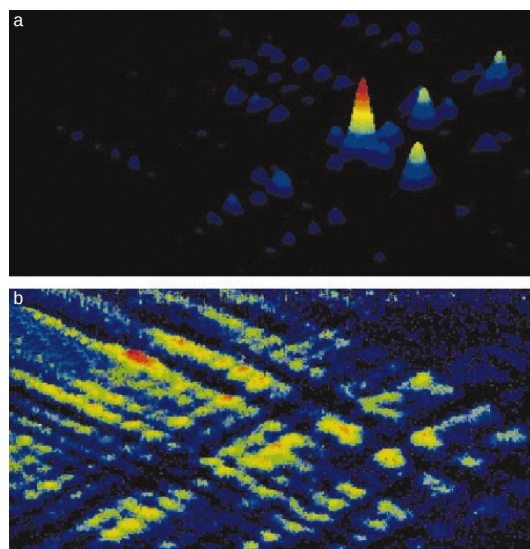


Figure 6 Experimental results in 2D photorefractive lattices. **a**, Observation of a two-dimensional discrete soliton in a biased highly nonlinear SBN:75 photorefractive crystal at a sufficiently high optical intensity. The power levels involved are in the milliwatt range. **b**, Discrete diffraction of an optical beam in a 2D waveguide lattice when the optical intensity is low⁴⁸. In both cases, the waveguide spacing (D) was 11 μm and the wavelength was 0.488 μm .

hexagonal and ‘diatomic’ lattices. Figure 6a demonstrates DS formation for a sufficiently high optical intensity, whereas Fig. 6b shows 2D discrete diffraction of an optical beam under linear conditions.

In addition, 2D DSs are possible in microstructured or photonic crystal fibres (PCFs), where a 2D waveguide lattice can be permanently embedded within the bulk. Very recently, nonlinear light localization has been experimentally observed in 2D optical fibre arrays⁴⁸. This is a promising technology because nonlinear interactions can take advantage of the long propagation lengths offered by silica fibres, thus reducing power requirements. The formation of self-localized states in active fibre laser arrays is also a topic of current interest^{49,50}.

In principle, DSs exist in a 3D environment^{45,46}. One such possibility may be in coupled-resonator optical waveguides (CROWs)⁵¹, where waveguiding is accomplished through light hopping among successive high- Q microcavities that effectively act like defects when embedded in a photonic crystal⁵². In this system, it is predicted that DSs can propagate without distortion along chains of coupled nonlinear microcavities. However, unlike their spatial cousins in waveguide arrays, DSs in this type of environment are by nature spatiotemporal entities⁴⁵. These states are possible as a result of the balance between discrete lattice dispersion and material nonlinearity. These self-localized entities can exhibit very low group velocities, depending on the coupling strength among successive microcavities and in some cases may remain immobile like frozen bubbles of light^{45,46}. In addition, this class of solitons can be effectively routed along any pre-assigned path in a 3D environment.

Soliton navigation and switching in waveguide array networks

One of the most important functions of a photonic network is to route information from a particular point of origin A to its final destination port Z . In such optical systems it is often highly desirable for routing to be accomplished completely optically so as to avoid unnecessary opto-electronic conversion. If, for example, data are re-directed by a space-switching matrix or a crossbar, it is cru-

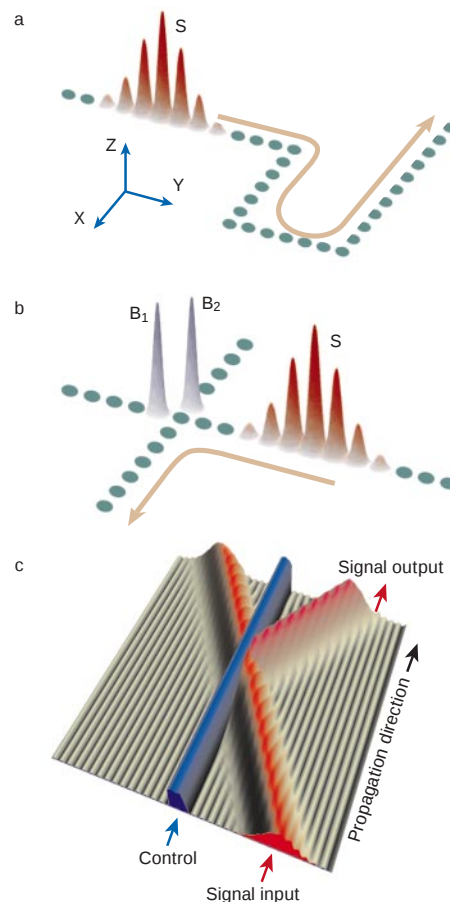


Figure 7 Possible applications of discrete solitons. **a**, A nonlinear array network involving consecutive bends. Light in this array propagates along the z -axis and is confined in the transverse x - y plane. The waveguide cross-sections are shown in green. A discrete soliton, S , shown in red, is set in motion in this system by appropriately tilting the beam. Computer simulations indicate that DSs can successfully negotiate a sequence of bends, producing very little radiation/reflection loss⁵⁴. The DS follows the pre-assigned path and remains essentially invariant during propagation. These losses can be effectively minimized by engineering the corner of the bend. **b**, An X-switching junction that uses two different DS families, the so-called ‘signals’ and ‘blockers’, denoted by S and B , respectively. Signals are moderately confined DSs, whereas blockers (depicted in blue) are strongly confined, occupying effectively one site. Unlike signals, which are highly mobile, blockers tend to retain their position after a collision event. In the junction shown, the blockers B_1 and B_2 interact ‘incoherently’ (different colours or polarizations) with S . The signal soliton S is routed towards the lower branch, because of the presence of the two blockers at the entries of the respective pathways. Had one of the two blockers not been present at the junction, the signal DS would have totally disintegrated into transmitted and reflected waves. Thus in essence, the junction operates as an AND gate⁵³. Animations illustrating these processes can be found at the website indicated in ref. 59. **c**, All-optical routing of an input signal beam to a specific output position using a parametric interaction in a quadratic nonlinear waveguide array. A tilted ($kD=\pi/2$) signal input beam crosses the array, producing virtually no diffraction. A control beam at about the second harmonic wavelength is used to generate, and all-optically deflect, an idler beam.

cial that this process occurs with minimum diffraction-induced crosstalk losses among nodes. DSs can offer a promising solution to this problem. By exploiting the nonlinearity of the array, discrete diffraction effects can be effectively counteracted, therefore eliminating undesirable crosstalk among sites (which serve as the network nodes). The success of such schemes, of course, relies on the

availability of fast-responding highly nonlinear materials. Even though waveguide arrays have thus far been realized in material systems that lack the required figures of merit (regarding their nonlinear properties), research into new materials promises to overcome these problems. In that case, one can envisage all-optical switching networks that play a vital role in tomorrow's communication systems.

Quite recently, it has been theoretically demonstrated that DSs in 2D nonlinear waveguide array networks can provide a rich environment for all-optical data processing applications⁵³. More specifically, this family of solitons can realize intelligent functional operations such as routing, blocking, logic functions and time-gating. These DSs can be routed anywhere in the network along pre-assigned array pathways that act like 'soliton wires' as explained in the proposed architectures of Fig. 7. Even more importantly, DSs can be routed at array intersections using vector-incoherent interactions with other DSs^{53–55}. In essence, these intersections behave as DS switching junctions. Such systems can also be used to realize logic operations (see Fig. 7b).

The distinct features of discreteness can also be exploited for other important applications in optical signal processing. Above, we have shown how nonlinearity and, in particular, DSs offer such possibilities. However, the engineering of linear wave diffraction is by itself highly promising in terms of controlling light in discrete configurations without the need for the power levels required in soliton-based schemes. In specifically designed configurations, discrete diffraction can be tailored so that signal beams suffer from virtually no diffraction-induced losses even at low power levels.

The dependence of the Bloch oscillation amplitude on the transverse index gradient is just one example of how waveguide arrays can be used to steer a single input beam to several output positions⁵⁶. An experimental realization of this concept demonstrated that a temperature change of only 5 K was sufficient to address six separate output channels.

The quadratic parametric interaction in waveguide arrays is another example of how light signals can be all-optically manipulated while complying with the low-power signal demands of modern telecommunication systems⁵⁷. In the switching configuration, which is shown in Fig. 7c, the input signal beam travels at an angle where diffraction is zero ($k_x D = \pi/2$). In this case, the control beam does not diffract because the stronger confinement in the waveguide renders the coupling zero. This allows two diffractionless beams to cross at an arbitrary position in the waveguide array. Parametric interaction provides phase-insensitive all-optical switching by generating a third diffractionless output beam. By implementing this scheme in the highly mature PPLN waveguide technology, routing applications of 100 Gbit/s signals should be feasible.

The field of optical dynamics in linear and nonlinear discrete waveguide lattices is at an exciting stage of development. Even though some of the basic concepts regarding discrete optical arrays have been around for a while, much remains to be understood and discovered. It is only very recently that fabrication techniques have matured to the point that experiments can be successfully conducted in these systems. This flurry of activity will allow us to assess the potential of these discrete structures for use in tomorrow's optical communication systems and switching networks. And, as is always the case in science, the best is yet to come. □

doi:10.1038/nature01936

1. Jones, A. L. Coupling of optical fibers and scattering in fibers. *J. Opt. Soc. Am.* **55**, 261–271 (1965).
2. Somekh, S., Garmire, E., Yariv, A., Garvin, H. L. & Hunsperger, R. G. Channel optical waveguide directional couplers. *Appl. Phys. Lett.* **22**, 46–48 (1973).
3. Christodoulides, D. N. & Joseph, R. I. Discrete self-focusing in nonlinear arrays of coupled waveguides. *Opt. Lett.* **13**, 794–796 (1988).
4. Davydov, A. S. Theory of contraction of proteins under their excitation. *J. Theor. Biol.* **38**, 559–569 (1973); Solitons and energy-transfer along protein molecules. *J. Theor. Biol.* **66**, 377–387 (1977).
5. Scott, A. C. & Macneil, L. Binding energy versus nonlinearity for a 'small' stationary soliton. *Phys. Lett. A* **98**, 87–88 (1983).
6. Lederer, F., Darmanyan, S. & Kobaykov, A. in *Spatial Optical Solitons* (eds Trillo, S. and Torruellas, W. E.) 269–292 (Springer-Verlag, New York, 2001).

7. Aceves, A. B. *et al.* Discrete self-trapping, soliton interactions, and beam steering in nonlinear waveguide arrays. *Phys. Rev. E* **53**, 1172–1189 (1996).
8. Krokolkowski, W. & Kivshar Yu, S. Soliton-based optical switching in waveguide arrays. *J. Opt. Soc. Amer. B* **13**, 876–887 (1996).
9. Kivshar Yu, S. Self-localization in arrays of defocusing waveguides. *Opt. Lett.* **18**, 1147–1149 (1993).
10. Eisenberg, H. S., Silberberg, Y., Morandotti, R., Boyd, A. R. & Aitchison, J. S. Discrete spatial optical solitons in waveguide arrays. *Phys. Rev. Lett.* **81**, 3383–3386 (1998).
11. Eisenberg, H. S., Silberberg, Y., Morandotti, R. & Aitchison, J. S. Diffraction management. *Phys. Rev. Lett.* **85**, 1863–1866 (2000).
12. Pertsch, T., Zentgraf, T., Peschel, U., Brauer, A. & Lederer, F. Anomalous refraction and diffraction in discrete optical systems. *Phys. Rev. Lett.* **88**, 093901 (2002).
13. Morandotti, R., Peschel, U., Aitchison, J. S., Eisenberg, H. S. & Silberberg, Y. Experimental observation of linear and nonlinear optical Bloch oscillations. *Phys. Rev. Lett.* **83**, 4756–4759 (1999).
14. Pertsch, T., Dannberg, P., Elflein, W., Brauer, A. & Lederer, F. Optical Bloch oscillations in temperature tuned waveguide arrays. *Phys. Rev. Lett.* **83**, 4752–4755 (1999).
15. Mandelik, D., Eisenberg, H. S., Silberberg, Y., Morandotti, R. & Aitchison, J. S. Band-gap structure of waveguide arrays and excitation of Floquet-Bloch solitons. *Phys. Rev. Lett.* **90**, 053902 (2003).
16. Pertsch, T. *et al.* Discrete solitons in quadratic nonlinear waveguide arrays, in *OSA Trends in Optics and Photonics* **80** (Optical Society of America, Washington, 2002).
17. Fleischer, J. W., Carmon, T., Segev, M., Efremidis, N. K. & Christodoulides, D. N. Observation of discrete solitons in optically-induced real time waveguide arrays. *Phys. Rev. Lett.* **90**, 023902 (2003).
18. Fleischer, J. W., Segev, M., Efremidis, N. K. & Christodoulides, D. N. Observation of two-dimensional discrete solitons in optically-induced nonlinear photonic lattices. *Nature* **422**, 147–150 (2003).
19. Yablonovitch, E. Inhibited spontaneous emission in solid-state physics and electronics. *Phys. Rev. Lett.* **58**, 2059–2062 (1987).
20. John, S. Strong localization of photons in certain disordered dielectric superlattices. *Phys. Rev. Lett.* **58**, 2486–2489 (1987).
21. Ioannopoulos, J. D., Villeneuve, P. R. & Fan, S. Photonic crystals: putting a new twist on light. *Nature* **386**, 143–149 (1997).
22. Russell, P. Photonic crystal fibers. *Science* **299**, 358–362 (2003).
23. Sievers, A. J. & Takeno, S. Intrinsic localized modes in anharmonic crystals. *Phys. Rev. Lett.* **61**, 970–973 (1988).
24. Dahan, M. B., Peik, E., Reichel, J., Castin, Y. & Salomon, C. Bloch oscillations of atoms in an optical potential. *Phys. Rev. Lett.* **76**, 4508–4511 (1996).
25. Anderson, B. P. & Kasevich, M. A. Macroscopic quantum interference from atomic tunnel arrays. *Science* **282**, 1686–1689 (1998).
26. Trombettoni, A. & Smerzi, A. Discrete solitons and breathers with dilute Bose-Einstein condensates. *Phys. Rev. Lett.* **86**, 2353–2356 (2001).
27. Bloch, F. Über die Quantenmechanik der Elektronen in Kristallgittern. *Z. Phys.* **52**, 555–600 (1928).
28. Zener, C. A theory of the electrical breakdown of solid dielectrics. *Proc. R. Soc. London Ser. A* **145**, 523–529 (1932).
29. Waschke, C. *et al.* Coherent submillimeter-wave emission from Bloch oscillations in a semiconductor superlattice. *Phys. Rev. Lett.* **70**, 3319–3322 (1993).
30. Peschel, U., Pertsch, T. & Lederer, F. Optical Bloch oscillations in waveguide arrays. *Opt. Lett.* **23**, 1701–1703 (1998).
31. Stegeman, G. I. & Segev, M. Optical spatial solitons and their interactions: universality and diversity. *Science* **286**, 1518–1522 (1999).
32. Ashcroft, N. & Mermin, N. *Solid State Physics* (Saunders, Philadelphia, 1976).
33. Ablowitz, M. J. & Clarkson, P. A. *Solitons, Nonlinear Evolution Equations and Inverse Scattering* (Cambridge Univ. Press, Cambridge, 1992).
34. Kivshar, Yu. S., Krokolkowski, W. & Chubykalo, O. A. Dark solitons in discrete lattices. *Phys. Rev. E* **50**, 5020–5032 (1994).
35. Morandotti, R., Eisenberg, H. S., Silberberg, Y., Sorel, M. & Aitchison, J. S. Self-focusing and defocusing in waveguide arrays. *Phys. Rev. Lett.* **86**, 3296–3299 (2001).
36. Morandotti, R., Peschel, U., Aitchison, J. S., Eisenberg, H. S. & Silberberg, Y. Dynamics of discrete solitons in optical waveguide arrays. *Phys. Rev. Lett.* **83**, 2726–2729 (1999).
37. Darmanyan, S., Kobaykov, A., Schmidt, E. & Lederer, F. Strongly localized vectorial modes in nonlinear waveguide arrays. *Phys. Rev. E* **57**, 3520–3530 (1998).
38. Ablowitz, M. J. & Musslimani, Z. H. Discrete diffraction managed spatial solitons. *Phys. Rev. Lett.* **87**, 254102–254104 (2001).
39. Sukhorukov, A. A. & Kivshar, Yu. S. Discrete gap solitons in modulated waveguide arrays. *Opt. Lett.* **27**, 2112–2114 (2002).
40. Bang, O. & Miller, P. D. Exploiting discreteness for switching in waveguide arrays. *Opt. Lett.* **21**, 1105–1107 (1996).
41. Meier, J. *et al.* Phase-control beam interactions in Kerr-nonlinear waveguide arrays, in *OSA Trends in Optics and Photonics* **80** (Optical Society of America, Washington, 2002).
42. Meier, J. *et al.* Discrete vector Kerr spatial solitons in AlGaAs array waveguides, in *OSA Trends in Optics and Photonics* **80** (Optical Society of America, Washington, 2002).
43. Peschel, T., Peschel, U. & Lederer, F. Discrete bright solitary waves in quadratically nonlinear media. *Phys. Rev. E* **57**, 1127–1133 (1998).
44. Malomed, B. & Kevrekidis, P. G. Discrete vortex solitons. *Phys. Rev. E* **64**, 26601 (2001).
45. Christodoulides, D. N. & Efremidis, N. K. Discrete temporal solitons along a chain of nonlinear coupled microcavities embedded in photonic crystals. *Opt. Lett.* **27**, 568–570 (2002).
46. Iliev, R., Peschel, U. & Lederer, F. Light propagation via coupled defects in photonic crystals. *Proc. Quantum Electronics Laser Science Conference* (2002).
47. Efremidis, N. K., Sears, S., Christodoulides, D. N., Fleischer, J. W. & Segev, M. Discrete solitons in photorefractive optically induced photonic lattices. *Phys. Rev. E* **66**, 046602 (2002).
48. Pertsch, T. *et al.* Light localization in disordered two-dimensional fiber arrays. *Phys. Rev. Lett.* (in the press).
49. Cheo, P. K., Liu, A. & King, G. G. A high-brightness laser beam from a phase-locked multicore Yb-doped fiber laser array. *IEEE Photon. Technol. Lett.* **13**, 439–441 (2001).
50. Efremidis, N. K. & Christodoulides, D. N. Discrete Ginzburg-Landau solitons. *Phys. Rev. E* **67**, 026606 (2003).
51. Yariv, A., Xu, Y., Lee, R. K. & Scherer, A. Coupled-resonator optical waveguide: a proposal and analysis.

- Opt. Lett.* **24**, 711–713 (1999).
52. Mingaleev, S. F., Kivshar, Yu. S. & Sammut, R. A. Long-range interaction and nonlinear localized modes in photonic crystal waveguides. *Phys. Rev. E* **62**, 5777–5782 (2000).
 53. Christodoulides, D. N. & Eugenueva, E. D. Blocking and routing discrete solitons in two-dimensional networks of nonlinear waveguide arrays. *Phys. Rev. Lett.* **87**, 233901 (2001).
 54. Christodoulides, D. N. & Eugenueva, E. D. Minimizing bending losses in two-dimensional discrete soliton networks. *Opt. Lett.* **26**, 1876–1878 (2001).
 55. Eugenueva, E. D., Efremidis, N. K. & Christodoulides, D. N. Design of switching junctions for two-dimensional discrete soliton networks. *Opt. Lett.* **26**, 1978–1980 (2001).
 56. Pertsch, T., Zentgraf, T., Peschel, U., Bräuer, A. & Lederer, F. Beam steering in waveguide arrays. *Appl. Phys. Lett.* **80**, 3247–3249 (2002).
 57. Pertsch, T., Peschel, U. & Lederer, F. All-optical switching in quadratically nonlinear waveguide arrays. *Optics Lett.* **28**, 102–104 (2003).
 58. Russell, P. St. J. Optics of Floquet-Bloch waves in dielectric gratings. *Appl. Phys. B* **39**, 231–246 (1986).
 59. Christodoulides, D. N. & Eugenueva, E. D. Animations of the processes described in Figs 7a, b as obtained after numerically solving the underlying evolution equations. AIP EPAPS: E-PRLTAO-87-018147 at <<http://www.aip.org/pubservs/epaps.html>>.

Acknowledgements The authors acknowledge useful discussions with Stewart Aitchison, George Stegeman, Mordechai Segev, Ulf Peschel, Roberto Morandotti, Hagai Eisenberg, Nikos Efremidis, Thomas Pertsch, D. Mandelik and Jared Hudock. The work of D.N.C. was supported by ARO MURI and of F.L. by the European Community grant (IST-2000-26005).

Surface plasmon subwavelength optics

William L. Barnes¹, Alain Dereux² & Thomas W. Ebbesen³

¹School of Physics, University of Exeter, EX4 4QL, UK (e-mail: w.l.barnes@ex.ac.uk)

²Laboratoire de Physique, Université de Bourgogne, BP 47870, F-21078 Dijon, France (adereux@u-bourgogne.fr)

³ISIS, Université Louis Pasteur, BP 70028, F-67083, Strasbourg Cedex, France (e-mail: ebbesen@isis-ulp.org)

Surface plasmons are waves that propagate along the surface of a conductor. By altering the structure of a metal's surface, the properties of surface plasmons—in particular their interaction with light—can be tailored, which offers the potential for developing new types of photonic device. This could lead to miniaturized photonic circuits with length scales that are much smaller than those currently achieved. Surface plasmons are being explored for their potential in subwavelength optics, data storage, light generation, microscopy and bio-photonics.

Surface plasmons (SPs) are of interest to a wide spectrum of scientists, ranging from physicists, chemists and materials scientists to biologists. Renewed interest in SPs comes from recent advances that allow metals to be structured and characterized on the nanometre scale. This in turn has enabled us to control SP properties to reveal new aspects of their underlying science and to tailor them for specific applications. For instance, SPs are being explored for their potential in optics, magneto-optic data storage, microscopy and solar cells, as well as being used to construct sensors for detecting biologically interesting molecules.

SPs were widely recognized in the field of surface science following the pioneering work of Ritchie in the 1950s (ref. 1). SPs are waves that propagate along the surface of a conductor, usually a metal. These are essentially light waves that are trapped on the surface because of their interaction with the free electrons of the conductor (strictly speaking, they should be called surface plasmon polaritons to reflect this hybrid nature²). In this interaction, the free electrons respond collectively by oscillating in resonance with the light wave. The resonant interaction between the surface charge oscillation and the electromagnetic field of the light constitutes the SP and gives rise to its unique properties.

For researchers in the field of optics, one of the most attractive aspects of SPs is the way in which they help us to concentrate and channel light using subwavelength structures. This could lead to miniaturized photonic circuits with length scales much smaller than those currently achieved^{3,4}. Such a circuit would first convert light into SPs, which would then propagate and be processed by logic elements, before being converted back into light. To build such a circuit one would require a variety of components: waveguides, switches, couplers and so on. Currently much effort is being devoted to developing such SP devices; one example is the 40 nm thick gold stripe that acts as a waveguide for SPs in Fig. 1. An appealing feature is that, when embedded in dielectric materials, the circuitry used to propagate SPs can also be used to carry electrical signals. Developments such as this raise the prospect of a new branch of photonics using SPs, sometimes called plasmonics.

The use of SPs to help us concentrate light in subwavelength structures stems from the different (relative) permittivities, ϵ , of the metals and the surrounding non-conducting media. (ϵ is the square of the complex index of refraction.)

Concentrating light in this way leads to an electric field enhancement that can be used to manipulate light-matter interactions and boost non-linear phenomena. For example, metallic structures much smaller than the wavelength of light are vital for the massive signal enhancement achieved in surface-enhanced Raman spectroscopy (SERS)—a technique that can now detect a single molecule^{5,6}. Furthermore, the enhanced field associated with SPs makes them suitable for use as sensors, and commercial systems have already been developed for sensing biomolecules. SP-based sensing applications and SERS will not be discussed further here because they are covered by other reviews^{7,8}.

Here we provide an overview of the properties of SPs and indicate why they are being considered for subwavelength optics. We examine how their propagation can be manipulated and discuss some of the optical components that have so far been demonstrated. We conclude by highlighting the practical potential of this field and indicate just how much work remains to be done for that potential to be realized.

Coupling to surface plasmons

The interaction between the surface charges and the electromagnetic field that constitutes the SP has two consequences (see Box 1). First, the interaction between the surface charge density and the electromagnetic field results in the momentum of the SP mode, $\hbar k_{\text{sp}}$, being greater than that of a free-space photon of the same frequency, $\hbar k_0$. ($k_0 = \omega/c$ is the free-space wavevector.) Solving Maxwell's equations under the appropriate boundary conditions yields the SP dispersion relation⁹, that is, the frequency-dependent SP wave-vector, k_{sp} ,

$$k_{\text{sp}} = k_0 \sqrt{\frac{\epsilon_d \epsilon_m}{\epsilon_d + \epsilon_m}} \quad (1)$$

The frequency-dependent permittivity of the metal, ϵ_m , and the dielectric material, ϵ_d , must have opposite signs if SPs are to be possible at such an interface. This condition is satisfied for metals because ϵ_m is both negative and complex (the latter corresponding to absorption in the metal). As an example, using equation (1), the SP wavevector for a silver-air interface in the red part of the visible spectrum is found to be $k_{\text{sp}} \approx 1.03 k_0$. This increase in momentum is associated with the binding of the SP to the surface, and the resulting momentum mismatch between light and SPs of the same frequency must be bridged if light is to be used to generate SPs.

The second consequence of the interaction between the surface charges and the electromagnetic field is that, in contrast to the propagating nature of SPs along the surface, the field perpendicular to the surface decays exponentially with distance from the surface. The field in this perpendicular direction is said to be evanescent or near field in nature and is a consequence of the bound, non-radiative nature of SPs, which prevents power from propagating away from the surface.

There are three main techniques by which the missing momentum can be provided. The first makes use of prism coupling to enhance the momentum of the incident light^{10,11}. The second involves scattering from a topological defect on the surface, such as a subwavelength protrusion or hole, which provides a convenient way to generate SPs locally^{3,12}. The third makes use of a periodic corrugation in the metal's surface¹³. Indeed, over 100 years ago, Wood¹⁴ reported anomalous behaviour in the diffraction of light by metallic diffraction gratings—some of these phenomena are now known to arise from coupling to SPs. The diffraction (scattering) of light by a metallic diffraction grating allows incident light to be momentum matched and thus coupled to SPs¹⁵. Importantly the reverse process also allows the otherwise non-radiative SP mode to couple with light in a controlled way with good efficiency^{16,17}, which is vital if SP-based photonic circuits are to be developed.

Surface plasmon propagation

Once light has been converted into an SP mode on a flat metal surface it will propagate but will gradually attenuate owing to losses arising from absorption in the metal. This attenuation depends on the dielectric function of the metal at the oscillation frequency of the SP. The propagation length, δ_{SP} , can be found by seeking the imaginary part, k''_{SP} , of the complex surface plasmon wavevector, $k_{SP} = k'_{SP} + ik''_{SP}$, from the SP dispersion equation (1) (ref. 15),

$$\delta_{SP} = \frac{1}{2k''_{SP}} = \frac{c}{\omega} \left(\frac{\epsilon'_m + \epsilon_d}{\epsilon'_m \epsilon_d} \right)^{\frac{3}{2}} \frac{(\epsilon'_m)^2}{\epsilon''_m} \quad (2)$$

where ϵ'_m and ϵ''_m are the real and imaginary parts of the dielectric function of the metal, that is, $\epsilon_m = \epsilon'_m + i\epsilon''_m$. Silver is the metal with the lowest losses in the visible spectrum: propagation distances are typically in the range 10–100 μm , increasing towards 1 mm as one moves into the 1.5 μm near-infrared telecom band (see Box 2). In the past, absorption by the metal was seen as such a significant problem that SPs were not considered viable for photonic elements; the SP propagation length was smaller than the size of components at that time. This view is now changing thanks primarily to recent demonstrations of SP-based components that are significantly smaller than the propagation length¹⁸. Such developments open the way to integrate many SP-based devices into circuits before propagation losses become too significant.

In addition to dealing with the problem of loss owing to absorption in the metal, there is another key loss mechanism that must be considered: unwanted coupling to radiation. To build SP-based circuits one will need components that convert one SP mode into another, for example, a switch to re-route SPs without scattering the SP mode in such a way as to lose its energy to freely propagating light.

Surface plasmon band structure and periodic surfaces

One of the key developments in photonics in the past 15 years has been that of photonic bandgap (PBG) materials. These synthetic materials use wavelength-scale periodic structures to manipulate the interaction between light and matter so as to build new photonic structures—a good example being that of the photonic crystal fibre¹⁹. These developments have been predominantly made in periodically structured insulating and semiconducting materials. By making use of SPs, metals too can be used as PBG materials, this time in the form of photonic surfaces²⁰.

The nature of SPs changes when they propagate on metal surfaces that are periodically textured on the scale of the wavelength of light.

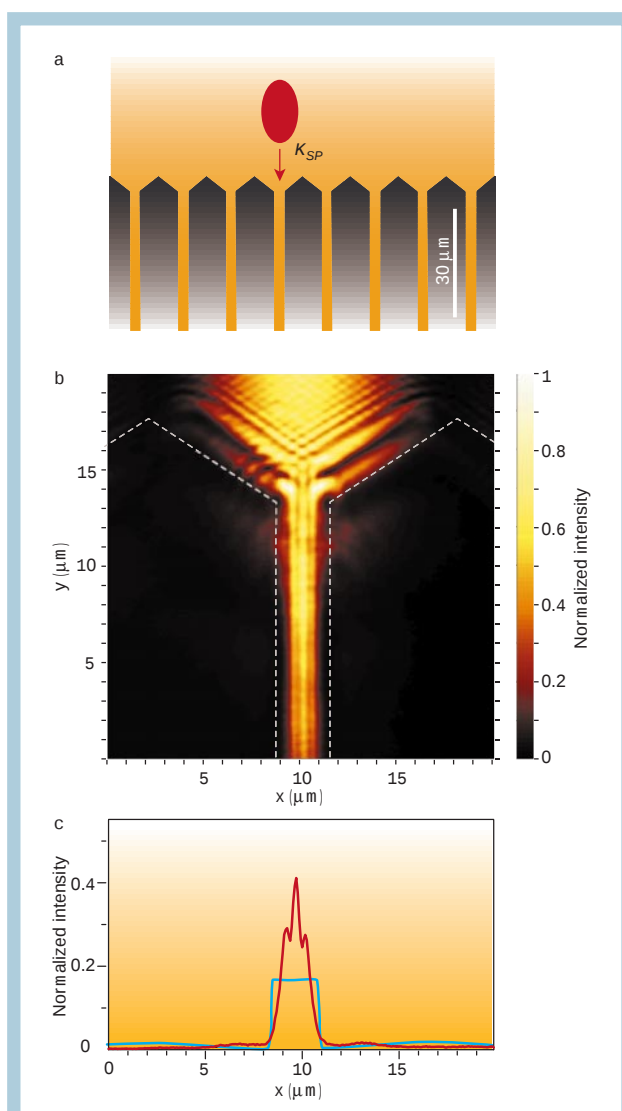


Figure 1 An SP waveguide. **a**, a scanning electron micrograph image of a 40 nm thick, 2.5 μm wide gold stripe lying on a glass substrate (courtesy of J. C. Weeber, Université de Bourgogne, France). **b**, The optical functionality of the stripe visualized by PSTM; an extended SP, launched on the larger area by a spot (indicated by the red ellipse) generated by total internal reflection illumination (wavelength=800 nm) through the substrate, is used to excite one of the stripe's SP eigen modes featuring three maxima. The photon scanning tunnelling microscopy (PSTM) image demonstrates that SPs are bound to the metal. **c**, A cross-section across the stripe shows that this mode is much better confined to the guiding material (indicated by the AFM topology—pale blue line) sustaining the mode than would be the case in dielectric-based waveguides. Not only the height of the guide but also the square root of the waveguide cross-section features a subwavelength size, underlining the fact that the SP mode is essentially bound to the metal surface rather than being a standing wave confined inside the metal volume.

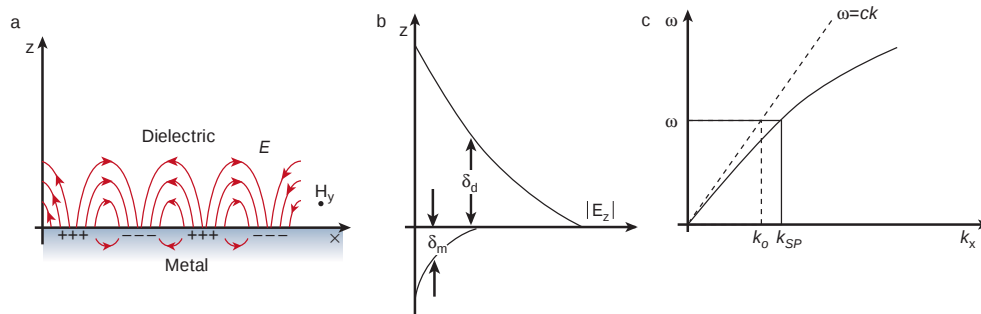
When the period of the nanostructure is half that of the effective wavelength of the SP mode, scattering may lead to the formation of SP standing waves and the opening of an SP stop band²¹ (see Box 3). When the surface is modulated in both in-plane directions, for example by a periodic array of bumps, SP modes may be prevented from travelling in any in-plane direction, thus leading to a full PBG for SP modes^{20,22} (Fig. 2). Fig. 2 provides a nice demonstration of how the problems associated with absorption by the metal can be overcome. It shows that, although finite, the SP propagation length is more than enough to allow the SP to

Box 1

Surface plasmon basics

SPs at the interface between a metal and a dielectric material have a combined electromagnetic wave and surface charge character as shown in **a**. They are transverse magnetic in character (**H** is in the y direction), and the generation of surface charge requires an electric

field normal to the surface. This combined character also leads to the field component perpendicular to the surface being enhanced near the surface and decaying exponentially with distance away from it (**b**). The field in this perpendicular direction is said to be evanescent, reflecting the bound, non-radiative nature of SPs, and prevents power from propagating away from the surface. In the dielectric medium above the metal, typically air or glass, the decay length of the field, δ_d , is of the order of half the wavelength of light involved, whereas the decay length into the metal, δ_m , is determined by the skin depth. **c**, The dispersion curve for a SP mode shows the momentum mismatch problem that must be overcome in order to couple light and SP modes together, with the SP mode always lying beyond the light line, that is, it has greater momentum (hk_{SP}) than a free space photon (hk_0) of the same frequency ω .



experience many periods of the textured surface and thus display the PBG phenomenon. We also note with interest that recent developments in the fabrication of periodic nanostructures via self-assembly offer the prospect of easily producing SP PBG substrates to act as photonic substrates on which to define SP photonic circuits.

At frequencies within a bandgap, the density of SP modes is zero—no SP modes can be supported. However, at the band edges, the SP mode dispersion is flat and the associated density of SP modes is high, corresponding to a high field enhancement close to the metal surface. Further, the nature of this flat band means that such modes can be excited by light that is incident over a wide range of angles, making them good candidates for frequency-selective surfaces. Flat bands are also associated with the localized SP modes of metallic nanoparti-

cles^{23,24}. The frequency and width of these modes are determined by the particle's shape, material, size and environment^{23,25,26}, and for this reason they are being pursued as tags for biosensing^{27,28} and as substrates for SERS²⁹ and potentially as aerials for fluorophores^{30,31}. The interaction between two or more nanoparticles can lead to still further levels of field enhancement^{32–34}, with even more dramatic effects associated with hot spots in random structures³⁵.

Mapping surface plasmons and developing components

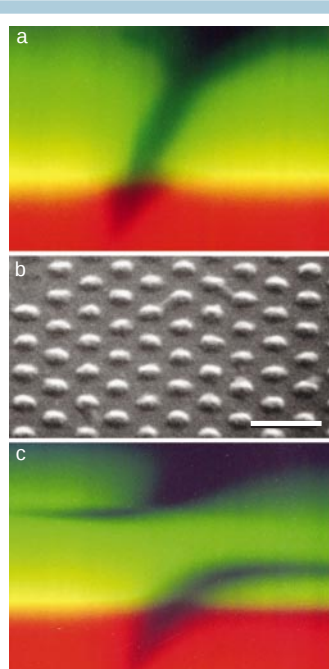
The properties of SP devices are intimately linked to the activity and the distribution of SPs on the metal surface. Much is still not known about the relationship between surface topology and the nature of the SP modes, and so a more detailed study of the details of this SP activity is vital. Because of the way SPs are confined to the surface and because of the subwavelength nature of the structures and fields involved, one cannot rely on traditional far-field techniques. Instead near-field techniques^{3,36} such as photon scanning tunnelling microscopy (PSTM) are typically employed to map the fields on the metal surface, for example, those of the SP waveguide in Fig. 1. A PSTM is basically a collection mode scanning near-field optical microscope where the sample lies on a glass prism, which enables one to shine light in total internal reflection. The nanometre size tip, mostly obtained by pulling an optical fibre, which may eventually be coated with a metal, frustrates the total reflection when scanning close to the surface and thereby maps the near-field intensities.

Surprisingly, as we enhance the capabilities of near-field techniques further to map the SP fields into the subwavelength regime we come up against an interesting variant of Heisenberg's uncertainty principle. Applied to the optical field, this principle says that we can only measure the electric (E) or the magnetic field (H) with accuracy when the volume δl^3 in which they are contained is significantly smaller than the wavelength of light in all three spatial dimensions. More precisely, Heisenberg's uncertainty principle binds E and H of the optical wave to δl through the cyclic permutation of their vector components (ij),

$$\Delta E_i \Delta H_j \geq \hbar c^2 / 2 \delta l^4. \quad (3)$$

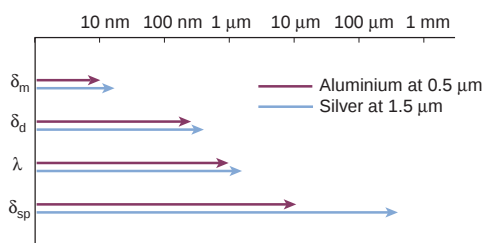
As volumes smaller than the wavelength are probed, measurements of optical energy become uncertain, highlighting the difficulty with performing measurements in this regime.

Figure 2 SP photonic bandgap. The SP dispersion curve shown in Box 1 was directly imaged using a modified prism coupling technique. **a**, The dispersion curve (here shown as inverse wavelength versus angle) for a flat surface is shown in the upper picture; here dark regions correspond to coupling of incident light to the SP mode, and the colours are produced on a photographic film by the wavelength of the light used. **b**, If the metal surface is textured with a two-dimensional pattern of bumps on an appropriate length scale (roughly half the wavelength of light) as shown in this SEM, a bandgap is introduced into the dispersion curve of the associated SP modes. Bar, 0.7 μm . **c**, The bandgap is clearly seen in the lower picture where there is a spectral region in which no SP mode (as indicated by the dark regions) exists. Also note the distortion of the SP mode and the edges of the bandgap.



Box 2

Surface plasmon length scales



There are three characteristic length scales that are important for SP-based photonics in addition to that of the associated light. The propagation length of the SP mode, δ_{sp} , is usually dictated by loss in the metal. For a relatively absorbing metal such as aluminium the propagation length 2 μm at a wavelength of 500 nm. For a low loss metal, for example, silver, at the same wavelength it is increased to 20 μm . By moving to a slightly longer wavelength, such as 1.55 μm , the propagation length is further increased towards 1 mm. The propagation length sets the upper size limit for any photonic circuit based on SPs. The decay length in the dielectric material, δ_d , is typically of the order of half the wavelength of light involved and dictates the maximum height of any individual features, and thus components, that might be used to control SPs. The ratio of $\delta_{sp}:\delta_d$ thus gives one measure of the number of SP-based components that may be integrated together. The decay length in the metal, δ_m , determines the minimum feature size that can be used; as shown in the diagram, this is between one and two orders of magnitude smaller than the wavelength involved, thus highlighting the need for good control of fabrication at the nanometre scale. The combinations chosen give an indication of range from poor (Al at 0.5 μm) to good (Ag at 1.5 μm) SP performance.

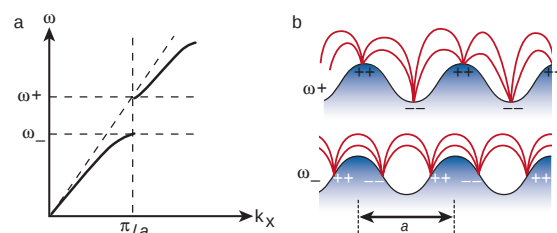
This is not as serious a limitation as one might imagine, because, although we would often like to measure the optical energy (for which we need to know both the electric and magnetic field strengths), many optical interactions are dominated by just one of the fields, and so mapping just one of them on a subwavelength scale can provide invaluable new insights. For example, most PSTM tips provide images that are proportional to the mapping of the distribution of $|E|^2$ in the near-field zone³⁶. Surprisingly, SPs with circular symmetry, sustained by a suitable metal coating of the tip, played a crucial role in the recent demonstration³⁷ that the distribution of $|H|^2$ associated with the optical wave can also be detected. Not only does this provide new possibilities for SPs, but it also enhances their role in such phenomena as the magneto-optic Kerr effect^{38,39} in extending the realm in which SPs have important photonic applications.

Near-field mapping techniques such as PSTM have been essential for the development of SP devices such as waveguides and other components. For example, they were used to map the field distribution associated with an SP waveguide based on a metal stripe^{40–42} that is illustrated in Fig. 1. Other strategies for waveguiding have also been explored, for example, using a well defined stripe defect on a periodically modulated photonic surface, with the defect acting as a guide⁴³. It was recently pointed out that SP waveguides can be obtained by exploiting roughness-induced Anderson localization, which inhibits SP propagation. In this case, waveguides are formed by channels flattened across an otherwise rough metal film⁴⁴. Waveguides can also be made of aligned metal (for example, gold) nanoparticles: a recent PSTM study has demonstrated the feasibility of laterally squeezing the optical near field by coupling localized SPs of an ensemble of linearly aligned gold nanoparticles⁴⁵, which suggests another approach to SP guiding⁴⁶.

By making use of advanced lithographic techniques to texture the

Box 3

Surface plasmon bandgaps



Periodic texturing of the metal surface can lead to the formation of an SP photonic bandgap when the period, a , is equal to half the wavelength of the SP, as shown in the dispersion diagram (a). Just as for electron waves in crystalline solids, there are two SP standing wave solutions, each with the same wavelength but, owing to their different field and surface charge distributions, they are of different frequencies. The upper frequency solution, ω_+ , is of higher energy because of the greater distance between the surface charges and the greater distortion of the field, as shown schematically in b. SP modes with frequencies between the two band edges, ω_+ and ω_- , cannot propagate, and so this frequency interval is known as a stop gap. By providing periodic texture in two dimensions, SP propagation in all in-plane directions can be blocked, leading to the full bandgap for SPs. At the band edges the density of SP states is high, and there is a significant increase in the associated field enhancement.

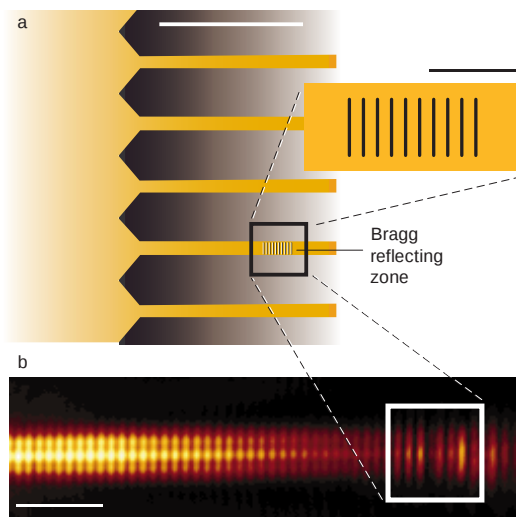


Figure 3 Surface plasmon Bragg reflector. **a**, An example of integration involving a Bragg reflector on a SP waveguide is provided by carving a series of slots with suitable sizes and separations in a gold stripe supported on a glass substrate, as shown on the SEM image (a) of a stripe similar to the one introduced in Fig. 1 (courtesy of E. Devaux, Université Louis Pasteur, France, and J. C. Weeber, Université de Bourgogne, France). White bar, 30 μm ; black bar, 2.5 μm . The incident SP, with three maxima (see Fig. 1), is excited by total internal reflection illumination as described in the legend of Fig. 1. The SP propagates from the left to the right until it reaches the Bragg reflecting zone (indicated by the blue squares). The mirror effect is clearly observed in the PSTM image (b) as the interference between incident and reflected SPs. To the right of the mirror zone, the intensity falls dramatically to a value not detectable by the PSTM tip. Bar, 2.5 μm .

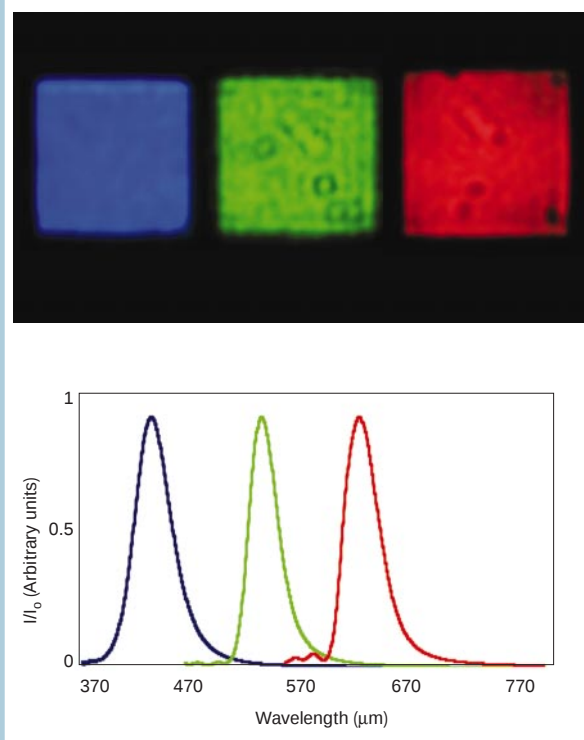


Figure 4 Normal incidence transmission for subwavelength holes. Normal incidence transmission images (top) and spectra (bottom) for three square arrays of subwavelength holes. For the blue, green and red arrays, the periods were 300, 450 and 550 nm, respectively, the hole diameters were 155, 180 and 225 nm and the peak transmission wavelengths 436, 538 and 627 nm. The arrays were made in a free standing 300 nm thick silver film (courtesy of A. Degiron, Université Louis Pasteur, France). Only the lowest order peak ($i, j = 0.1$ in equation (4)) of the spectrum of each array is shown as it dominates the colour seen. The figure shows that nanostructures can control the resonant wavelength of SP phenomena.

metal surface with protrusions, slits or holes, lenses and mirrors for SPs may be integrated into circuits not only on an extended thin film¹⁸ but also directly on a finite-width stripe. An example of such a structure is shown in Fig. 3, where a Bragg reflector for SP modes is demonstrated. Although not yet fully optimized, these early demonstrations strengthen the prospect for SP-based photonic elements.

Hole arrays

The demonstration of enhanced transmission through periodic arrays of subwavelength holes in optically thick metallic films has sparked new interest in SPs^{47–50}. Not only is the transmission much higher than expected from classic diffraction theory, it can be greater than the percentage area occupied by the holes, implying that even the light impinging on the metal between the holes can be transmitted. In other words, the whole periodic structure acts like an antenna in the optical regime, nicely demonstrating the benefits that SP modes can provide. The transmission spectra of hole arrays display peaks that can be tuned by adjusting the period and the symmetry as shown in Fig. 4. For a square array of period a_0 , the peaks λ_{max} in the normal incidence transmittance spectra can be identified approximately from the dispersion relation (equation (1)), and they are given by:

$$\lambda_{\text{max}} \sqrt{i^2 + j^2} \approx a_0 \sqrt{\frac{\epsilon_m \epsilon_d}{\epsilon_m + \epsilon_d}} \quad (4)$$

where indices i and j are the scattering orders from the array⁴⁸.

For a free-standing hole array, incident light is diffracted/scattered by the array, producing evanescent waves that tunnel through the holes, resulting in a small but finite amplitude on the far side of the array. Here the evanescent waves are again diffracted/scattered; the interference of the resulting waves produces the light that propagates away from the structure. Equation (4) acts as a starting point in analysing the transmission spectrum, a spectrum that is more accurately determined by taking into account these diffraction/interference effects^{51–55}. SPs act to enhance the fields associated with the evanescent waves, thus producing a way to increase the transmittance. When the metal film is thin enough, this tunneling may become resonant because the SP modes on the two surfaces can overlap and interact via the holes^{49–51}. Interestingly, photon entanglement can be preserved or lost upon transmission through a hole, depending on the experimental conditions⁵⁶.

Single apertures

Individual subwavelength apertures can also exhibit enhanced transmission when surrounded by a periodic structure that harvests the incident light. If such nanostructure is added to the output side of the film then, remarkably, the emerging light can be beamed rather than diffracted. In addition, the light is mostly re-emitted from a very small area surrounding the aperture⁵⁷. This suggests that a well directed source of light can be generated using a subwavelength aperture, an exciting development that is being pursued as a source for a variety of optical technologies, including high-density magneto-optic data storage. A theoretical treatment⁵⁸ shows that just a few well designed features around the exit of a slit aperture can give rise to beaming, as confirmed recently in the microwave region⁵⁹ and shown vividly in the numerical simulation presented in Fig. 5.

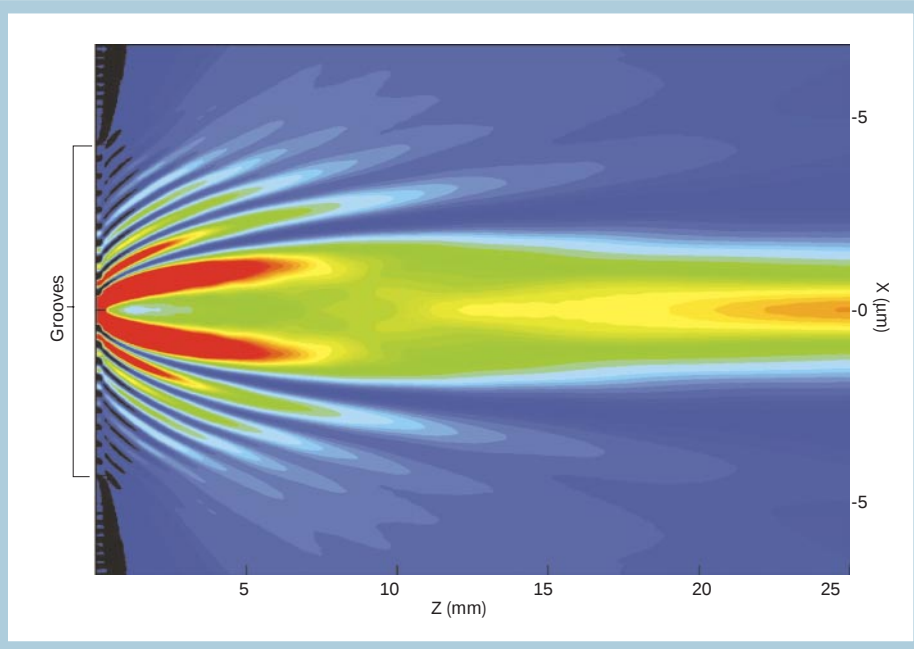
In the single hole and hole array structures described above, the holes do not support any propagating modes, that is, the diameter of the hole is smaller than $\lambda/2$ (half the wavelength of the light that is transmitted), and tunnelling is necessarily part of the transmission mechanism. However, when the aperture is larger than $\lambda/2$ in at least one dimension, such as a slit, then the aperture can support propagating modes^{60–62}; the slit can even be wrapped up in the form of an annulus⁶³. For instance, investigations of the transmission spectrum of periodic slit arrays reveal that both the slit modes and the modes caused by the periodic modulation of the surface (standing wave surface modes) play a role. In the case of a single slit surrounded by linear grooves, the transmission spectrum is dominated by the modes caused by the periodic groove structure, although other modes play an important role⁶⁴.

Future directions

In addition to their use in photonic circuits, SPs are being explored for other photonic technologies, notably the generation of light. A particular example is that of the organic light-emitting diode. Here excitons are produced by injection of charge into a semi-conducting organic layer that is typically only 100 nm thick. The close proximity of the excitons to the metallic cathode used to inject electrons means that much of the power that would otherwise have been radiated is in fact lost to SP modes on the cathode surface^{65–68}, thus detracting from device efficiency. Managing and, if necessary, recovering this power through the use of a periodic nanostructure is likely to be important for many future device designs, especially for high-efficiency, long-life and full-colour devices. The inverse process, that of using SPs to enhance the absorption of light, for example, in solar cells^{69,70}, is also of interest. SP modes have even been employed as the lasing mode in quantum cascade lasers⁷¹.

To develop SP-based photonics, non-linear components such as switches will also be required. The unique properties of SPs give them some advantages in this respect. We mentioned that they are characterized by an evanescent, near field that is enhanced close to the surface. The use of SPs to enhance non-linear processes such as harmonic generation is well known^{72–74}, but it is only recently that experiments have been designed to look at how non-linear effects may be used to provide

Figure 5 Calculated pattern of light emerging from a single slit surrounded by a finite array of grooves (courtesy of F. J. García-Vidal, Universidad Autónoma de Madrid, Spain) and L. Martín-Moreno (Universidad de Zaragoza, Spain). For this simulation there were 10 grooves on each side of the slit, each of width 40 nm and depth 100 nm. The groove period was 500 nm. The contour plot shows the spatial dependence of the component of the Poynting vector along the radial direction for a wavelength of 560 nm, corresponding to a spectral transmission maximum. Blue indicates low intensity (3×10^{-4}), and red high (1).



new elements for SP-based photonics⁷⁵. This is an area in urgent need of further work.

SPs show great potential as a new class of subwavelength photonic devices; their particular attractions are their near-field character and their associated field enhancements. Nanoscale lithography enables researchers to explore the underlying science and has already led to proof-of-concept demonstrations, notably of waveguides, reflectors, beam splitters, enhanced transmission and beaming. Some of these components may find their way into the world of photonics as discrete elements. The greater dream of making a completely plasmonic circuit will require much further research. It is an exciting challenge that is stimulating activity around the world. □

doi:10.1038/nature01937

- Ritchie, R. H. Plasma losses by fast electrons in thin films. *Phys. Rev.* **106**, 874–881 (1957).
- Burstein, E. in *Polaritons* (eds Burstein, E. & De Martini, F.) 1–4 (Pergamon, New York, 1974).
- Hecht, B., Bielefeldt, H., Novotny, L., Inoué, Y. & Pohl, D. W. Local excitation, scattering, and interference of surface plasmons. *Phys. Rev. Lett.* **77**, 1889–1892 (1996).
- Pendry, J. Playing tricks with light. *Science* **285**, 1687–1688 (1999).
- Kneipp, K. *et al.* Single molecule detection using surface-enhanced Raman scattering (SERS). *Phys. Rev. Lett.* **78**, 1667–1670 (1997).
- Nie, S. M. & Emery, S. R. Probing single molecules and single nanoparticles by surface-enhanced Raman scattering. *Science* **275**, 1102–1106 (1997).
- Homola, J., Yee, S. S. & Gauglitz, G. Surface plasmon resonance sensors: review. *Sensors Actuat. B* **54**, 3–15 (1999).
- Kneipp, K., Kneipp, H., Itzkan, I., Dasari, R. R. & Feld, M. S. Surface enhanced Raman scattering and biophysics. *J. Phys. C* **14**, R597–R624 (2002).
- Sambles, J. R., Bradbery, G. W. & Yang, F. Z. Optical-excitation of surface-plasmons - an introduction. *Contemp. Phys.* **32**, 173–183 (1991).
- Kretschmann, E. & Raether, H. Radiative decay of nonradiative surface plasmons excited by light. *Z. Naturforsch.* **A23**, 2135–2136 (1968).
- Otto, A. Excitation of nonradiative surface plasma waves in silver by the method of frustrated total reflection. *Z. Phys.* **216**, 398 (1968).
- Ditlbacher, H. *et al.* Fluorescence imaging of surface plasmon fields. *Appl. Phys. Lett.* **80**, 404–406 (2002).
- Ritchie, R. H., Arakawa, E. T., Cowan, J. J. & Hamm, R. N. Surface-plasmon resonance effect in grating diffraction. *Phys. Rev. Lett.* **21**, 1530–1533 (1968).
- Wood, R. W. On a remarkable case of uneven distribution of light in a diffraction grating spectrum. *Phil. Mag.* **4**, 396 (1902).
- Raether, H. *Surface Plasmons* (ed. Hohl, G.) (Springer, Berlin, 1988).
- Moreland, J., Adams, A. & Hansma, P. K. Efficiency of light emission from surface plasmons. *Phys. Rev. B* **25**, 2297–2300 (1982).
- Worthing, P. T. & Barnes, W. L. Efficient coupling of surface plasmon polaritons to radiation using a bi-grating. *Appl. Phys. Lett.* **79**, 3035–3037 (2001).
- Ditlbacher, H., Krenn, J. R., Schider, G., Leitner, A. & Aussenegg, F. R. Two-dimensional optics with surface plasmon polaritons. *Appl. Phys. Lett.* **81**, 1762–1764 (2002).
- Cregan, R. F. *et al.* Single-mode photonic band gap guidance of light in air. *Science* **285**, 1537–1539 (1999).

- Kitson, S. C., Barnes, W. L. & Sambles, J. R. A full photonic band gap for surface modes in the visible. *Phys. Rev. Lett.* **77**, 2670–2673 (1996).
- Barnes, W. L., Preist, T. W., Kitson, S. C. & Sambles, J. R. Physical origin of photonic energy gaps in the propagation of surface plasmons on gratings. *Phys. Rev. B* **54**, 6227–6244 (1996).
- Barnes, W. L., Kitson, S. C., Preist, T. W. & Sambles, J. R. Photonic surfaces for surface plasmon polaritons. *J. Opt. Soc. Am. A* **14**, 1654–1661 (1997).
- Kreibig, U. & Vollmer, M. *Optical properties of metal clusters* (Springer, Berlin, 1995).
- Special issue: Optical Properties of Nanoparticles. *Appl. Phys. B* **73** (2001).
- Haynes, C. L. & Van Duyne, R. P. Nanosphere lithography: A versatile nanofabrication tool for studies of size-dependent nanoparticle optics. *J. Phys. Chem. B* **105**, 5599–5611 (2001).
- Sonnichsen, C. *et al.* Spectroscopy of single metallic nanoparticles using total internal reflection microscopy. *Appl. Phys. Lett.* **77**, 2949–2951 (2000).
- Schultz, D. A. Plasmon resonant particles for biological detection. *Curr. Opin. Biotechnol.* **14**, 13–22 (2003).
- Oldenburg, S. J., Genick, C. C., Clark, K. A. & Schultz, D. A. Base pair mismatch recognition using plasmon resonant particle labels. *Anal. Biochem.* **309**, 109–116 (2002).
- Feld, J. N. *et al.* Optimized surface-enhanced Raman scattering on gold nanoparticle arrays. *Appl. Phys. Lett.* **82**, 3095–3097 (2003).
- Levi, S. A. *et al.* Fluorescence of dyes adsorbed on highly organized, nanostructured gold surfaces. *Chem.-a Eur. J.* **8**, 3808–3814 (2002).
- Vargas-Baca, I. *et al.* Linear and nonlinear optical responses of a dye anchored to gold nanoparticles dispersed in liquid and polymeric matrices. *Can. J. Chem.* **80**, 1625–1633 (2002).
- Rechberger, W. *et al.* Optical properties of two interacting gold nanoparticles. *Opt. Commun.* **220**, 137–141 (2003).
- Kottmann, J. P. & Martin, O. J. F. Retardation-induced plasmon resonances in coupled nanoparticles. *Opt. Lett.* **26**, 1096–1098 (2001).
- García-Vidal, F. J. & Pendry, J. B. Collective theory for surface enhanced Raman scattering. *Phys. Rev. Lett.* **77**, 1163–1166 (1996).
- Gresillon, S. *et al.* Experimental observation of localized optical excitations in random metal-dielectric films. *Phys. Rev. Lett.* **82**, 4520–4523 (1999).
- Weeber, J. C. *et al.* Observation of light confinement effects with a near-field optical microscope. *Phys. Rev. Lett.* **77**, 5332–5335 (1996).
- Devaux, E. *et al.* Local detection of the optical magnetic field in the near zone of dielectric samples. *Phys. Rev. B* **62**, 10504–10514 (2000).
- Silva, T. J. & Schultz, S. A scanning near-field optical microscope for the imaging of magnetic domains in reflection. *Rev. Sci. Instr.* **67**, 715–725 (1996).
- Hermann, C. *et al.* Surface-enhanced magneto-optics in metallic multilayer films. *Phys. Rev. B* **64**, 235422 (2001).
- Kottmann, J. P., Martin, O. J. F., Smith, D. R. & Schultz, S. Plasmon resonances of silver nanowires with a nonregular cross section. *Phys. Rev. B* **64**, 5402 (2001).
- Weeber, J. C., Dereux, A., Girard, C., Krenn, J. R. & Goudonnet, J. P. Plasmon polaritons of metallic nanowires for controlling submicron propagation of light. *Phys. Rev. B* **60**, 9061–9068 (1999).
- Weeber, J. C. *et al.* Near-field observation of surface plasmon polariton propagation on thin metal stripes. *Phys. Rev. B* **64**, 045411 (2001).
- Bozhevolnyi, S. I., Erland, J., Leosson, K., Skovgaard, P. M. W. & Hvam, J. M. Waveguiding in surface plasmon polariton band gap structures. *Phys. Rev. Lett.* **86**, 3008–3011 (2001).
- Bozhevolnyi, S. I., Volkov, V. S. & Leosson, K. Localization and waveguiding of surface plasmon polaritons in random nanostructures. *Phys. Rev. Lett.* **89**, 186801 (2002).
- Krenn, J. R. *et al.* Squeezing the optical near-field zone by plasmon coupling of metallic nanoparticles. *Phys. Rev. Lett.* **82**, 2590–2593 (1999).
- Maier, S. A. *et al.* Local detection of electromagnetic energy transport below the diffraction limit in metal nanoparticle plasmon waveguides. *Nat. Mat.* **2**, 229–232 (2003).

47. Ebbesen, T. W., Lezec, H. J., Ghaemi, H. F., Thio, T. & Wolff, P. A. Extraordinary optical transmission through sub-wavelength hole arrays. *Nature* **391**, 667–669 (1998).
48. Ghaemi, H. F., Thio, T., Grupp, D. E., Ebbesen, T. W. & Lezec, H. J. Surface plasmons enhance optical transmission through subwavelength holes. *Phys. Rev. B* **58**, 6779–6782 (1998).
49. Krishnan, A. *et al.* Evanescently coupled resonance in surface plasmon enhanced transmission. *Opt. Commun.* **200**, 1–7 (2001).
50. Degiron, A., Lezec, H. J., Barnes, W. L. & Ebbesen, T. W. Effects of hole depth on enhanced light transmission through subwavelength hole arrays. *Appl. Phys. Lett.* **81**, 4327–4329 (2002).
51. Martín-Moreno, L. *et al.* Theory of extraordinary optical transmission through subwavelength hole arrays. *Phys. Rev. Lett.* **86**, 1114–1117 (2001).
52. Bonod, N., Enoch, S., Li, L., Popov, E. & Nevière, M. Resonant optical transmission through thin metallic films with and without holes. *Opt. Expr.* **11**, 482–490 (2003).
53. Vigoureux, V. M. Analysis of the Ebbesen experiment in the light of evanescent short range diffraction. *Opt. Commun.* **198**, 257–263 (2001).
54. Salomon, L., Grillot, F. D., Zayats, A. V. & de Fornel, F. Near-field distribution of optical transmission of periodic subwavelength holes in a metal film. *Phys. Rev. Lett.* **86**, 1110–1113 (2001).
55. Darmanyan, S. A. & Zayats, A. V. Light tunneling via resonant surface plasmon polariton states and the enhanced transmission of periodically nanostructured metal films. *Phys. Rev. B* **67**, 035424 (2003).
56. Altwischer, E., van Exter, M. P. & Woerdman, J. P. Plasmon-assisted transmission of entangled photons. *Nature* **418**, 304–306 (2002).
57. Lezec, H. J. *et al.* Beaming light from a subwavelength aperture. *Science* **107**, 1895 (2002).
58. Martín-Moreno, L., García-Vidal, F. J., Lezec, H. J., Degiron, A. & Ebbesen, T. W. Theory of highly directional emission from a single sub-wavelength aperture surrounded by surface corrugations. *Phys. Rev. Lett.* **90**, 167401 (2003).
59. Hibbins, A. P., Sambles, J. R. & Lawrence, C. R. Gratingless enhanced microwave transmission through a subwavelength aperture in a thick metal plate. *Appl. Phys. Lett.* **81**, 4661–4663 (2002).
60. Porto, J. A., García-Vidal, F. J. & Pendry, J. B. Transmission resonances on metallic gratings with very narrow slits. *Phys. Rev. Lett.* **83**, 2845–2848 (1999).
61. García-Vidal, F. J. & Martín-Moreno, L. Transmission and focusing of light in one-dimensional periodically nanostructured metals. *Phys. Rev. B* **66**, 155412 (2002).
62. Barbara, A., Quémenerais, P., Bustarret, E. & Lopez-Rios, T. Optical transmission through subwavelength metallic gratings. *Phys. Rev. B* **66**, 161403 (2002).
63. Baida, F. I. & Van Labeke, D. Light transmission by subwavelength annular aperture arrays in metallic films. *Opt. Commun.* **209**, 17–22 (2002).
64. García-Vidal, F. J., Lezec, H. J., Ebbesen, T. W. & Martín-Moreno, L. Multiple paths to enhance optical transmission through a single subwavelength slit. *Phys. Rev. Lett.* **90**, 213901 (2003).
65. Weber, W. H. & Eagen, C. F. Energy transfer from an excited dye molecule to the surface plasmons of an adjacent metal. *Opt. Lett.* **4**, 236–238 (1979).
66. Pockrand, I., Brillante, A. & Möbius, D. Nonradiative decay of excited molecules near a metal surface. *Chem. Phys. Lett.* **69**, 499–504 (1980).
67. Barnes, W. L. Fluorescence near interfaces: the role of photonic mode density. *J. Mod. Opt.* **45**, 661–699 (1998).
68. Hobson, P. A., Wedge, S., Wasey, J. A. E., Sage, I. & Barnes, W. L. Surface plasmon mediated emission from organic light emitting diodes. *Adv. Mat.* **14**, 1393–1396 (2002).
69. Andrew, P., Kitson, S. C. & Barnes, W. L. Surface plasmon energy gaps and photoabsorption. *J. Mod. Opt.* **44**, 395–406 (1997).
70. Westphalen, M., Kreibig, U., Rostalski, J., Luth, H. & Meissner, D. Metal cluster enhanced organic solar cells. *Sol. Energy Mat. Sol. Cells* **61**, 97–105 (2000).
71. Tredicucci, A. *et al.* Single-mode surface-plasmon laser. *Appl. Phys. Lett.* **76**, 2164–2166 (2000).
72. Coutaz, J. L., Nevier, M., Pic, E. & Reinisch, R. Experimental study of surface-enhanced second-harmonic generation on silver gratings. *Phys. Rev. B* **32**, 2227–2232 (1985).
73. Quail, J. C. & Simon, H. J. Second-harmonic generation with phase-matched long-range and short-range surface plasmons. *J. Appl. Phys.* **56**, 2589–2591 (1984).
74. Tsang, T. Y. F. Surface-plasmon-enhanced third-harmonic generation in thin silver films. *Opt. Lett.* **21**, 245–247 (1996).
75. Smolyaninov, I. I., Zayats, A. V., Gungor, A. & Davis, C. C. Single-photon tunneling via localized surface plasmons. *Phys. Rev. Lett.* **88**, 187402 (2002).

Recent developments in compact ultrafast lasers

Ursula Keller

Physics Department, Swiss Federal Institute of Technology (ETH), Zürich, Switzerland (e-mail: keller@phys.ethz.ch)

Ultrafast lasers, which generate optical pulses in the picosecond and femtosecond range, have progressed over the past decade from complicated and specialized laboratory systems to compact, reliable instruments. Semiconductor lasers for optical pumping and fast optical saturable absorbers, based on either semiconductor devices or the optical nonlinear Kerr effect, have dramatically improved these lasers and opened up new frontiers for applications with extremely short temporal resolution (much smaller than 10 fs), extremely high peak optical intensities (greater than 10 TW/cm²) and extremely fast pulse repetition rates (greater than 100 GHz).

Six years after the first laser was demonstrated, De Maria and co-workers produced the first ultrashort pulses, estimated to be just picoseconds long, using a passively modelocked Nd:glass laser¹. Modelocking is a technique that generates ultrashort pulses from lasers. We distinguish between active and passive modelocking, and the latter can generate much shorter pulses using saturable absorbers.

The result from De Maria *et al.*¹ had an underlying problem: it did not measure a regular pulse train as described in Box 1. They actually measured an amplitude-modulated pulse train (Fig. 1). In this regime, called Q-switched modelocking, the modelocked picosecond pulses are modulated with a much longer Q-switched pulse envelope, which occurs at a much lower repetition rate. Lack of a regular pulse train remained a problem for passively modelocked solid-state lasers for more than 20 years.

Eventually the problem was solved in 1992 by using saturable absorbers from semiconductors². Since their introduction, the pulse durations, average powers, pulse energies and pulse repetition rates of compact ultrafast solid-state lasers have improved by several orders of magnitude through the addition of a saturable semiconductor cavity laser mirror (generally referred to as semiconductor saturable absorber mirrors, SESAMs)^{2–4}.

Compact efficient semiconductor diode lasers replaced inefficient lamps for optical excitation. Such diode-pumped solid-state lasers have further enhanced reliability and produced unsurpassed performance. For ultrabroadband tunability and to produce even shorter pulses, a faster, saturable absorber mechanism, referred to as Kerr lens modelocking⁵, is required. Currently, pulse durations can range from picoseconds to a few femtoseconds, depending on the laser material and the parameters of the saturable absorber. The average power has been increased to 60 W directly from a modelocked diode-pumped laser with pulse energies larger than 1 mJ⁶, and the pulse repetition rate has been increased to more than 100 GHz⁷. These improvements in performance of ultrashort lasers keep advancing the frontiers of ultrashort pulse durations, high average power and high pulse repetition rates, which will be discussed in more detail in this review. But first we continue with an overview of significant applications and a more detailed explanation of the two key enabling technologies, Kerr lens modelocking and semiconductor saturable absorbers. At this point, I believe that the development of ultrafast lasers has not come to its end but will continue to deliver superior performance for many applications.

Applications of ultrafast lasers

Ultrafast laser technology has developed rapidly over the past decade, and as our understanding develops so will the application of these lasers. By examining four features of ultrafast lasers, I will highlight some of the applications that benefit from laser development.

Ultrashort pulse duration

Ultrashort pulse duration allows fast temporal resolution. In the same way that a strobe light at a disco ‘freezes’ the motion of dancers, a modelocked laser can ‘freeze’ the motion of fast-moving objects such as molecules or electrons and therefore can measure the relaxation processes of carriers in semiconductors⁸, chemical reaction dynamics^{9,10} and electro-optical sampling of high-speed electronics^{11,12}. Using modelocked lasers, the dissociation dynamics of molecules and more complex chemical reaction dynamics have been measured, and this work was rewarded with a Nobel prize in chemistry for A. H. Zewail in 1999.

High pulse repetition rate

Lasers with multi-gigahertz repetition rates are key components of many applications. They are used in high-capacity telecommunication systems^{13,14}, photonic switching devices¹⁵, optical interconnections and for clock distribution. And applications of the future such as clocks for very large scale integrated (VLSI) microprocessors¹⁶, polarized electron beams¹⁷ for electron accelerators and high-speed electro-optic sampling techniques^{11,12} will rely on multi-gigahertz pulse trains with short pulses, low timing jitter and low amplitude noise.

As data transmission rates increase, modelocked lasers with a tunable wavelength of around 1.55 μm will become important in telecommunications. Transmission systems at 10 GHz and higher often use return-to-zero (RZ) pulse formats and soliton dispersion management techniques^{14,18}. These approaches benefit from the availability of simple, compact, transform-limited optical pulse generators^{14,20}. For example, they eliminate the need for a modulator to create the pulses and thereby simplify system architecture, increase efficiency and reduce cost. Additionally, pulse quality is typically very good—much better than with modulated CW (continuous wave) sources. This improves system signal-to-noise ratios and allows scaling to higher repetition rates through optical time-division multiplexing. Transmission of data at 160 Gbits per second through standard single-mode fibres has been demonstrated using such laser sources.

High average power 10–100 GHz sources at shorter wavelengths (for example, 1 μm and below) are promising sources

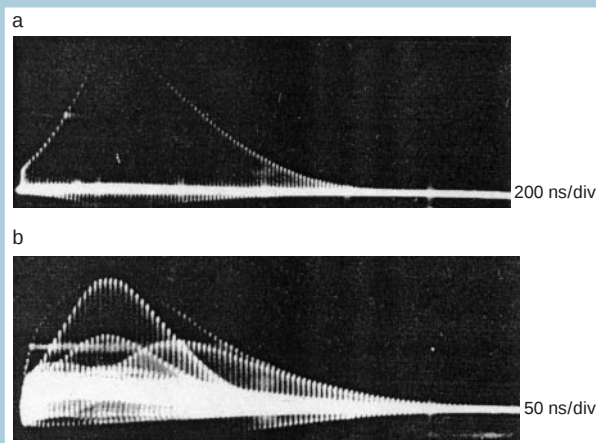


Figure 1 Pulses from a passively modelocked Nd:glass laser with a dye saturable absorber inside the laser resonator. Picosecond pulses were superimposed upon much longer pulses, which is referred to as Q-switched modelocking. Reprinted with permission from ref. 1 © American Institute of Physics.

for optical clocks in integrated circuits. Current microprocessor clocks in personal computers operate at more than 3 GHz rates, increasing between 15% to 30% annually. The International Technology Roadmap for Semiconductors 2001 (ITRS) forecasts predict a 40 GHz clock rate in 2020. Optical clocks, produced for example by a modelocked laser, can be precisely injected into specific circuits inside a VLSI microprocessor and have the potential to reduce on-chip power requirements, skew, jitter, and support scaling to high clock rates beyond 40 GHz.

Broad spectrum

A broad spectrum supports good spatial resolution for optical coherence tomography (OCT), a technique for non-invasive cross-sectional imaging in biological systems^{21,22}. OCT uses low coherence interferometry to produce a two-dimensional (2D) image of optical scattering from internal tissue microstructures; this process is similar to that of ultrasonic pulse-echo imaging. OCT does not necessarily need short pulses, but modelocked femtosecond lasers, alone or with additional external spectral broadening, offer a much higher average power than any other broadband source. Thus, these lasers produce a longitudinal and lateral spatial resolution of a few micrometres.

Pulse trains not only provide a broad spectrum but also a stable comb-shaped optical spectrum, where the spacing of the individual longitudinal modes exactly equals the pulse repetition rate (Box 1). Therefore, these lasers are stable multi-wavelength sources^{23,24}. The optical comb can be locked to the International Telecommunication Union grid, resulting in a novel multi-wavelength source that is suitable for dense wavelength-division-multiplexing (DWDM) applications²⁵.

More recently, the broadband frequency comb has been used for high precision optical frequency metrology^{26–28}. The frequency comb produces a 'ruler' in the frequency domain with which an unknown optical frequency can be measured. The frequency of the beat (the difference in frequency between the unknown frequency and the closest frequency in the ruler) is always smaller than one half of the comb period. Thus, an optical frequency in the 100 THz regime can be scaled down to a microwave frequency in the range of 1 GHz, which can be measured very accurately. This approach can be used for an all-optical atomic clock that is expected to outperform today's state-of-the-art caesium clocks²⁹.

These broadband frequency combs can also be used to stabilize the electric field underneath the pulse envelope^{30,31}, which is important in highly nonlinear processes such as photoionization³² and

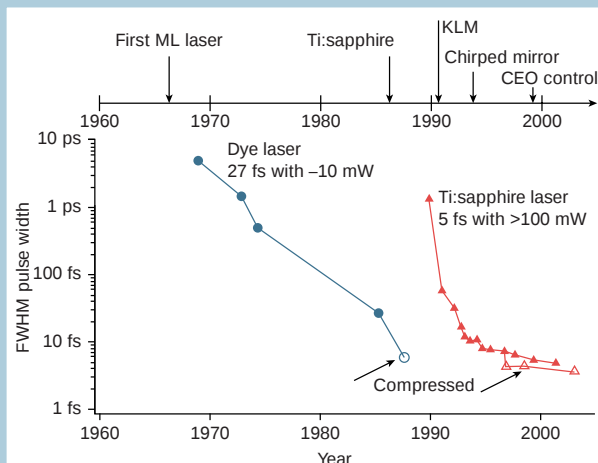


Figure 2 Improvements in ultrashort pulse generation since the first demonstration of a laser in 1960. Until the end of the 1980s, ultrashort pulse generation was dominated by dye lasers, and pulses as short as 27 fs with an average power of ≈ 10 mW were achieved at a centre wavelength of 630 nm (ref. 61). External pulse compression ultimately resulted in pulses as short as 6 fs—a world record result by C. V. Shank's group that was not surpassed for about 10 years⁵³. This situation changed with the discovery of the Ti:sapphire laser. Today, pulses with only two optical cycles at FWHM (full-width half-maximum) at a centre wavelength of 800 nm have been generated with Ti:sapphire lasers with more than 100 mW average output power^{62,72}. External compression resulted in pulses as short as 3.8 fs (ref. 76). The filled symbols indicate results directly achieved from a laser; open symbols indicate results achieved with additional external pulse compression.

high-harmonic generation³³. Such stabilized pulses can support single attosecond pulses in the soft X-ray spectral region^{33,34}.

High peak intensity

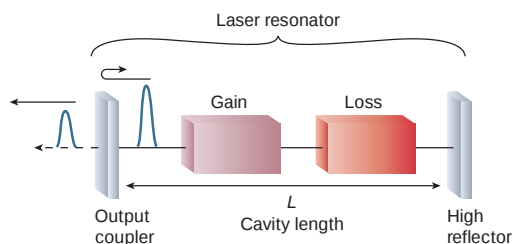
The high peak intensity of the pulse can be used to alter materials by 'cold' ablation (when a material is changed to gas directly from a solid) or to generate other colours/wavelengths through nonlinear frequency conversion. Diode-pumped solid-state lasers with high average power have produced femtosecond pulses with a pulse energy larger than 1 μ J with a 30 MHz pulse repetition rate. This is an unprecedented combination of short pulse, high pulse energy and high average power.

Such high peak intensity sources make 'non-thermal' ablation (without an increase in temperature) possible without any further amplification. The ability of intense ultrashort-pulse lasers to fabricate microstructures in solid targets is very promising, and the quality of ablated holes and patterns is much better using femtosecond or picosecond pulses instead of nanosecond pulses^{35–37}. Ultrafast 'non-thermal' melting in semiconductors has been observed, where a solid-to-liquid phase transition occurs faster than carrier-lattice equilibration times^{38,39}. This fast mode of material modification is important because it reduces undesirable effects caused by heat conduction and interaction of the pulse with ablated material. In medical applications, such short laser pulses have been used to obtain a higher level of surgical cutting precision, particularly in corneal surgery⁴⁰ and brain tumour removal⁴¹. These lasers reduce secondary damage effects, such as shock waves and cavitation bubbles, in tissues⁴² because the fluence threshold for optical breakdown decreases with pulse duration⁴³.

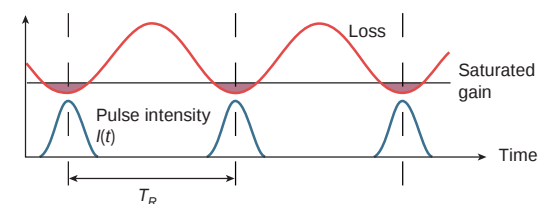
Simple external pulse compression combined with novel high average power solid-state lasers, at the full pulse repetition rate, results in peak powers as high as 12 MW with 33 fs pulses of the laser oscillator⁴⁴. This pulse can be focused to a peak intensity of 10^{14} W/cm², a level at which high-field laser physics, such as high harmonic generation^{45,46}

Box 1 Modelocking

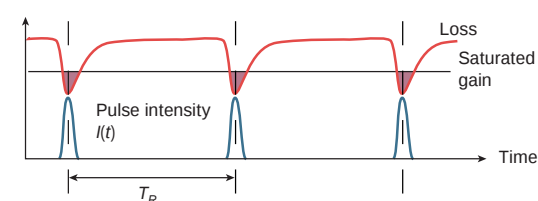
Modelocking is used to generate ultrashort pulses from lasers. A schematic set-up with a gain and a loss element inside a laser resonator is shown below. An output coupler partially transmits a small fraction of the laser pulse out of the laser resonator equally spaced by the resonator round-trip time. Typically an intracavity loss modulator is used to collect the laser light in short pulses around the minimum of the loss modulation with a period given by the cavity round-trip time $T_R = 2L/v_g$, where L is the laser cavity length and v_g the group velocity (that is, the propagation velocity of the peak of the pulse intensity). There are two type of modelocking: passive and active.



Active modelocking

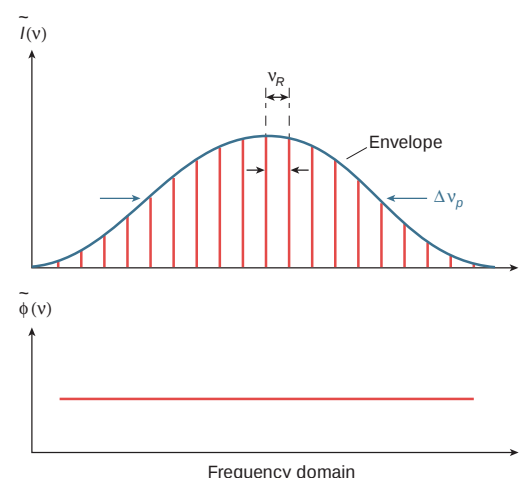
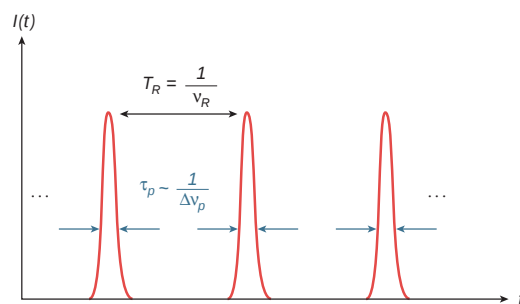


Passive modelocking



For active modelocking, an external signal is applied to an optical loss modulator typically using the acousto-optic or electro-optic effect. Such an electronically driven loss modulation produces a sinusoidal loss modulation with a period given by the cavity round-trip time T_R . The saturated gain at steady state then only supports net gain around the minimum of the loss modulation and therefore only supports pulses that are significantly shorter than the cavity round trip time.

For passive modelocking, a saturable absorber is used to obtain a self-amplitude modulation of the light inside the laser cavity. Such an absorber introduces some loss to the intracavity laser radiation, which is relatively large for low intensities but significantly smaller for a short pulse with high intensity. Thus, a short pulse then produces a loss modulation because the high intensity at the peak of the pulse saturates the absorber more strongly than its low intensity wings. This results in a loss modulation with a fast initial loss saturation (that is, reduction of the loss) determined by the pulse duration and typically a somewhat slower recovery that depends on the detailed mechanism of the absorption process in the saturable absorber. In effect, the circulating pulse saturates the laser gain to a level that is just sufficient to compensate for the losses from pulse itself, although any other circulating low-intensity light experiences more loss than gain and thus dies out during the following cavity round-trips. The obvious remaining question is how does passive modelocking start. Ideally, it starts from normal noise fluctuations in the laser. One noise spike is strong enough to significantly reduce its loss in the saturable absorber and thus will be more strongly amplified during the following cavity round trips, so that the stronger noise spike continues to further reduce its



loss and continues its growth until reaching steady state, where a stable pulse train has been formed.

Generally, we can obtain much shorter pulses with passive modelocking using a saturable absorber, because the recovery time of the saturable absorber can be very fast, resulting in a fast loss modulation. Modelocked pulses are much shorter than the cavity round-trip time and therefore can produce an ideal fast loss modulation that is inversely proportional to the pulse envelope. In comparison, any electronically driven loss modulation is significantly slower because of its sinusoidal loss modulation.

In the time domain, this means that a modelocked laser produces an equidistant pulse train, with a period defined by the round-trip time of a pulse inside the laser cavity T_R and a pulse duration τ_p . In the frequency domain, this results in a phase-locked frequency comb with a constant mode spacing that is equal to the pulse repetition rate $\nu_R = 1/T_R$. The spectral width of the envelope of this frequency comb is inversely proportional to the pulse duration.

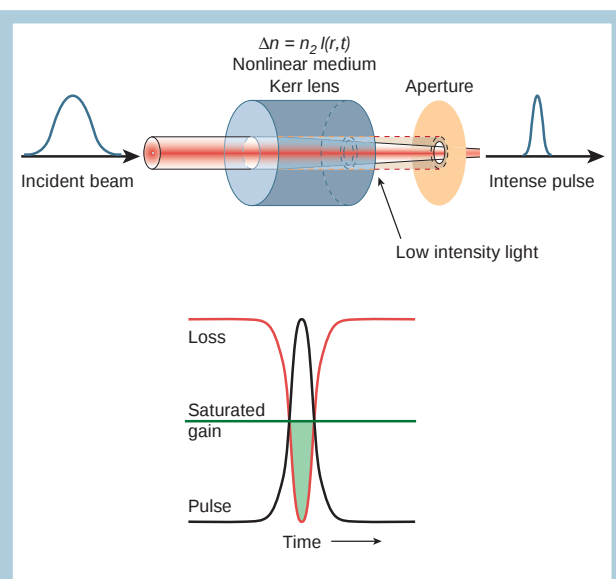


Figure 3 Kerr lens modelocking is obtained through a Kerr lens at an intracavity focus in the gain medium or in another material, where the refractive index increases with intensity $\Delta n = n_2 I(r, t)$, where n_2 is the nonlinear refractive index and $I(r, t)$ the radial- and time-dependent intensity of a short-pulsed laser beam. In combination with a hard aperture inside the cavity, the cavity design is made such that the Kerr lens reduces the laser mode area for high intensities at the aperture and therefore forms an effective fast saturable absorber. In most cases, however, soft-aperture KLM is used, where the reduced mode area in the gain medium improves for a short time the overlap with the (strongly focused) pump beam and therefore the effective gain. A significant change in mode size is only achieved by operating the laser cavity near one of the stability limits of the cavity.

and laser plasma generated X-rays⁴⁷, are possible. This increased pulse repetition rate would improve signal-to-noise ratios by four orders of magnitude compared with standard sources at kilohertz repetition rates. This is important for low-power applications such as X-ray imaging and microscopy⁴⁸, femtosecond extreme UV and soft X-ray photoelectron spectroscopy⁴⁹ and ultrafast X-ray diffraction^{38,39}.

I have outlined a few examples of ultrafast lasers, but many other applications already exist and many more will be developed in the future. Many of these applications are only possible because of the development of solid-state lasers and new modelocking techniques, particularly Kerr lens and semiconductor saturable absorber modelocking, which will be discussed next.

Ti:sapphire laser and Kerr lens modelocking (KLM)

The success of passively modelocked dye lasers in the 70s and 80s^{50,51} (Fig. 2) was partially due to the absence of the Q-switching problem present in solid-state lasers. These dye lasers were the first to generate sub-picosecond pulses⁵² and were the workhorses of research laboratories for time-resolved spectroscopy. External pulse compression resulted in 6 fs pulses⁵—a record that lasted for 10 years.

In the late 80s, Ti:sapphire was a newly discovered solid-state laser material that had the necessary broad gain bandwidth to support femtosecond pulses⁵⁴. By the end of the 80s, it was generally assumed that everything about modelocked Ti:sapphire lasers was understood, but there was a surprise to come. In 1990, two important papers were presented at consecutive conferences. First, Ishida *et al.*⁵⁵ presented a passively modelocked Ti:sapphire laser with an intracavity saturable absorber dye that produced stable 190 fs pulses. Second, Sibbett's group presented 60 fs pulses from a Ti:sapphire laser that appeared not to have a saturable absorber⁵⁶. This second result—in the absence of a visible saturable absorber—had an instant impact on the research community, but the ultrafast laser experts realized that

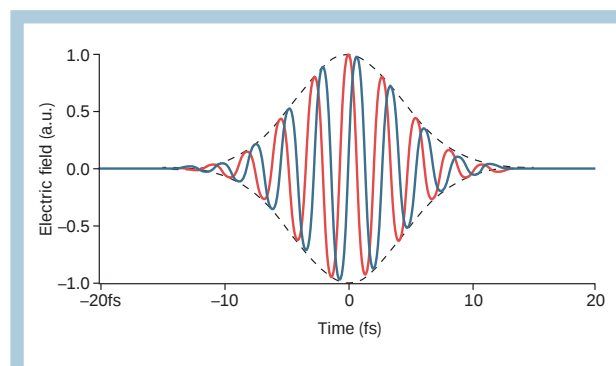


Figure 4 The electric field (carrier) and pulse envelope (dashed line) of a few-cycle pulse for different values of the CEO phase of 0 (red line) and $\pi/2$ (blue line).

the first result was also very surprising, even though a saturable absorber was present. It was clear that the dye saturable absorber, with a recovery time in the nanosecond range, could not support ultrashort pulses with a Ti:sapphire laser as it could with dye lasers. Sibbett's modelocking approach^{5,56} was initially termed 'magic modelocking', and it triggered a major research effort into understanding passive modelocking of solid-state lasers. This form of modelocking was soon explained^{57–59} and is now referred to as Kerr lens modelocking (KLM) (Fig. 3). Ishida's result was also explained by KLM: the slow dye saturable absorber only provided a reliable starting mechanism for KLM⁶⁰.

KLM allowed the community to shorten pulses to the few femtosecond regime. Because the Kerr lens produces a 'non-resonant' saturable absorber, it is inherently broadband, broader than any other saturable absorber available today. With ultrafast dye lasers, pulses as short as 27 fs with around 10 mW average power were generated⁶¹, but pulses around 5–6 fs with around 100 mW average power can be produced with Ti:sapphire lasers^{62,63}. In addition, very broad tunability with shorter than 100 fs pulses became possible for the first time. For time-resolved spectroscopy, these features had a major impact, because, for the first time, such lasers could be used on a wide range of materials.

KLM has some significant drawbacks. Generally, KLM lasers are not self-starting, that is, pulse formation does not start on its own, and additional perturbation to briefly increase laser noise is required. This is typically achieved by mechanically 'shaking' one of the laser cavity mirrors. Restrictive cavity designs^{64,65} allow partial self-starting, but this generally requires submillimetre precision for cavity mirror alignments, a very clean environment to minimize intracavity losses and a laser cavity operated close to the stability limit. In addition, optimization of self-starting conflicts with achieving the shortest possible pulses, because, when the cavity is optimized, the Kerr effect is too strong for very short pulses to react sensitively to long pulses during the start-up phase. In the picosecond regime, KLM lasers tend to be less stable because the Kerr lens becomes too weak. Therefore, the search for alternative solutions for compact ultrafast lasers continued.

Semiconductor saturable absorbers (SESAMs)

The breakthrough in the development of ultrafast lasers came with the semiconductor saturable absorber² (Box 2). In contrast to KLM lasers, the saturable absorber can be optimized independently of cavity design, allowing successful modelocking to be achieved with a broad range of solid-state lasers and cavity designs.

As early as 1974, a CO₂ laser⁶⁶, and in 1980 and 1984, a semiconductor diode laser⁶⁷ and a colour center laser⁶⁸, respectively, were passively modelocked with a semiconductor saturable absorber. In all these cases, the underlying physics of the modelocking process was significantly different from modelocking ion-doped solid-state

Box 2

Saturable absorbers and SESAMs

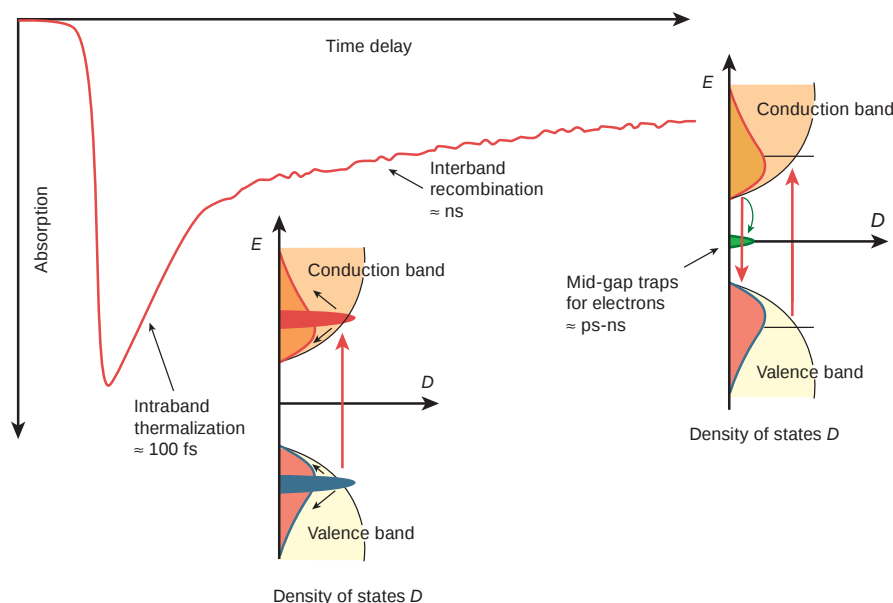
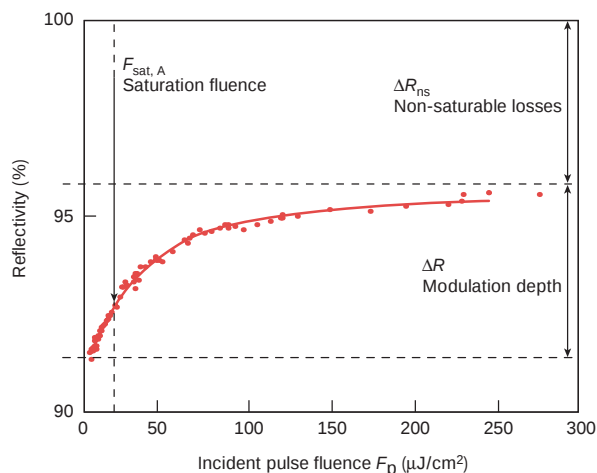
A saturable absorber is a material that has decreasing light absorption with increasing light intensity. We need saturable absorbers that show this effect at the intensities typically found in solid-state laser cavities, and semiconductor saturable absorbers are ideally suited for this. The key parameters for a saturable absorber are its wavelength range (where it absorbs), its dynamic response (how fast it recovers), and its saturation intensity and fluence (at what intensity or pulse energy density it saturates).

Saturable absorbers used in the past were typically dyes, which have short lifetimes, high toxicity and complicated handling procedures. Alternative solid-state saturable absorbers include crystals such as Cr:YAG, which typically operate for only a limited range of wavelengths, recovery times and saturation levels.

Semiconductor materials, however, can absorb over a broad range of wavelengths (from the visible to the mid-infrared). We can also control their absorption recovery time and saturation fluence (typically 1 to 100 $\mu\text{J}/\text{cm}^2$) by altering the growth parameters and device design.

Semiconductor saturable absorbers are integrated into a mirror structure, resulting in a device that reflects more light the more intense the light is. We call this general class of device semiconductor saturable absorber mirrors—SESAMs. The SESAM is a saturable absorber that operates in reflection, thus the reflectivity increases with higher incoming pulses (see the figure on the right). During the past decade we have significantly improved the device design, fabrication process and long-term device reliability. We have SESAM designs that can cover wavelengths from <800 nm to >1600 nm, pulsedwidths from femtoseconds to nanoseconds, and power levels from milliwatts to >100 watts.

A semiconductor absorbs light when the photon energy is sufficient to excite carriers from the valence band to the conduction band. Under conditions of strong excitation, the absorption is saturated because possible initial states of the pump transition are depleted while the final states are partially occupied. Within 60–300 fs of excitation, the carriers in each band thermalize, and this leads to a partial recovery of the absorption. On a longer time scale—typically between a few picoseconds and a few nanoseconds—the carrier will be removed by recombination and trapping. The presence of two different time scales can be rather useful for modelocking. The longer time constant results in a reduced saturation intensity for a part of the absorption, which facilitates self-starting modelocking, whereas the faster time constant is more effective in shaping subpicosecond pulses. Therefore, SESAMs allow us to easily obtain self-starting modelocking.



lasers such as Ti:sapphire, Nd:YAG, Yb:YAG, Cr:LiSAF and so on. The main difference is that these lasers have an emission cross-section that is typically more than 1,000 times smaller, and an upper state lifetime of the laser transition more than 1,000 times longer, than dye, diode and colour center lasers. The smaller the emission cross-section, the stronger the tendency for self-Q-switching (Box 3). This meant that the parameters of the saturable absorber had to be carefully adapted to prevent Q-switching instabilities, which partially explains why it took so long for the first successful demonstration. In addition, the theoretical background of Q-switched modelocking instabilities had not been fully worked out. Once this was under-

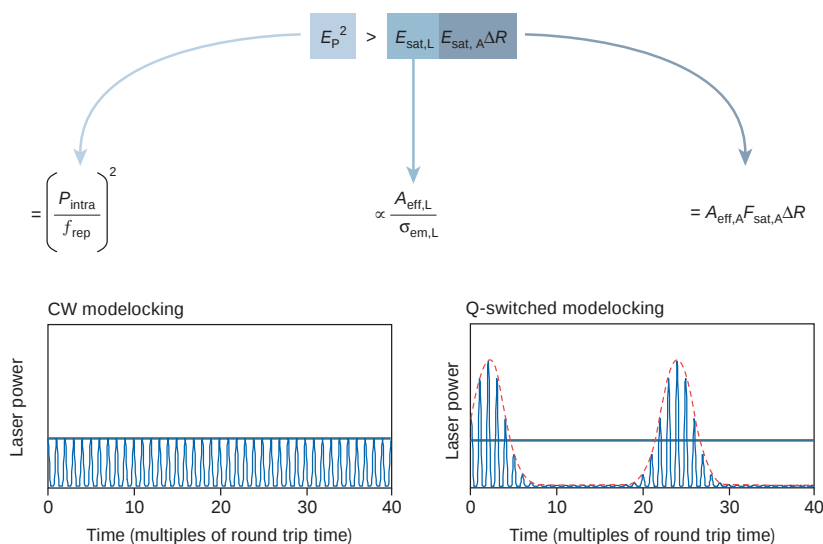
stood⁶⁹, much better performance in terms of average output power and high pulse repetition rates was ultimately achieved.

Saturable absorbers with the parameters, such as modulation depth, saturation fluence and nonsaturable losses (Box 2), required for a given laser cavity and pump power were obtained using semiconductor mirror devices, where one or more semiconductor saturable absorber layers are embedded inside a mirror structure. Different semiconductor materials provide a wide range of bandgaps for operation from the visible to the far-infrared spectral range, therefore, bandgap engineering can provide broad tunability. In addition, defect engineering for recovery times ranging from the nanosecond to the

Box 3

Q-switched modelocking

A saturable absorber may also produce Q-switched modelocking, where the laser emits bunches of modelocked pulses, which may or may not have a stable Q-switching envelope. Q-switching instabilities occur when the pulse energy is temporarily increased because of noise fluctuations in the laser, which then gets even further increased because of the stronger saturation of the saturable absorber. This has to be balanced by a stronger saturation of the gain. If the gain is not sufficiently saturated then the pulse energy will increase further and self-Q-switching occurs. The physical background of Q-switching instabilities can be summarized by a simple guideline: Q-switching instabilities in ion-doped solid-state lasers can be prevented if $E_p^2 > E_{\text{sat,L}} E_{\text{sat,A}} \Delta R$, assuming that the saturable absorber is fully saturated.



The important parameters are E_p intracavity energy, P_{intra} intracavity average power, f_{rep} pulse repetition rate, $E_{\text{sat,L}}$ saturation energy of the laser medium, $A_{\text{eff,L}}$ average mode area inside the laser medium, $\sigma_{\text{em,L}}$ emission cross-section of gain material, $E_{\text{sat,A}}$ saturation energy of the absorber, $A_{\text{eff,A}}$ average mode area inside the saturable absorber, $F_{\text{sat,A}}$ saturation fluence of absorber, ΔR modulation depth of the saturable absorber.

For femtosecond pulses, modelocking can be further supported by soliton formation, which relaxes the critical intracavity pulse energy by a factor of three to 10, depending on the specific parameters.

femtosecond regime is possible. Saturable absorbers can be adjusted to the required absorber parameter even when the mode size is fixed.

Typically the SESAM is operated with an incident pulse fluence of about three to five times that of the saturation fluence. This saturation level of the absorber provides nearly the maximum modulation depth without damaging the device⁷⁰. Higher saturation also reduces the tendency for Q-switching instabilities because of thermal effects and two-photon absorption, which becomes more significant for femtosecond pulses. The recovery time of the SESAM can be as long as 10 to 30 times relative to the final pulse duration⁷⁰, and with this form of modelocking, pulses as short as 13 fs have been generated⁷¹.

The success of semiconductor saturable absorbers for passively modelocked solid-state lasers results from their small saturation fluence and the additional benefits of cavity mirror device integration, sophisticated bandgap and defect engineering and epitaxial wafer-scale fabrication, which reduces its production cost substantially. SESAM design depends strongly on the laser parameters and can be optimized for different operational regimes. On the basis of KLM and SESAMs, the performance frontiers of pulsed solid-state lasers have been pushed forward by orders of magnitude during the past decade. These fast moving frontiers in pulse duration, average power and pulse repetition rate will be discussed next.

Frontiers: ultrashort pulses

At present, nothing is better than the Ti:sapphire laser as an ultrafast laser. Only KLM Ti:sapphire lasers have generated less than 6 fs pulse durations^{62,72}. At a centre wavelength of 800 nm, an optical cycle lasts only 2.7 fs; therefore, a pulse duration of 5.4 fs has only two optical cycles at full-width half-maximum of the pulse intensity. Significantly shorter pulses have only been generated using pulse compression in noble-gas-filled hollow fibres⁷³ or with synchronously pumped optical parametric oscillators⁷⁴. At the moment, the shortest pulses that are properly and fully characterized by an indepen-

dent phase measurement⁷⁵ in the visible and near-infrared spectral regime are 3.8 fs long⁷⁶. In this regime, the measurement of pulse duration is nearly as difficult as generating them. Unfortunately, new record pulses have been claimed without them being properly measured, because, generally, less careful measurement results in a 'shorter pulse'.

Dispersion management is an increasingly hard task in the few-cycle regime. Generally, when a pulse propagates through a medium it acquires a frequency-dependent phase shift and broadens. Dispersive pulse broadening is a linear effect and therefore can be compensated for by rearranging the different spectral components in time at a different place inside the laser cavity. Non-absorbing materials generally exhibit positive dispersion, and regrouping the different spectral components requires the opposite frequency-dependent phase shift (that is, negative dispersion).

Negative dispersion with low losses is typically obtained using geometrical effects, such as inserting a pair of prisms into the cavity⁷⁸, where different wavelength components will travel on slightly different optical paths. The negative dispersion obtained from this effect is proportional to the prism separation, and a positive dispersion results from propagation in the prism, but this is easily adjusted by altering how much of the prism is inserted into the optical path. However, residual higher-order dispersion normally limits the pulse duration in Ti:sapphire lasers to about 10 fs using fused quartz prism pairs^{79–81}.

Dielectric Bragg mirrors with regular quarter-wave stacks of alternating materials with high and low refractive indices have a negligibly small dispersion when operated within their reflection bandwidth but an increasing dispersion at the edges of this range. Modified designs can be used to obtain well controlled dispersion over a large wavelength range with chirped mirrors⁸², which give a much broader reflection bandwidth than any other mirror. With chirped mirrors, less than 10 fs pulses have been generated with good quality for the first time⁸³, and further improvements in chirped mirror designs, such as double

chirped mirrors^{84,85} and back-side coated chirped mirrors⁸⁶, have resulted in record pulsewidth results⁸⁷.

Ultrashort pulses in the one-to-two optical cycle regime also open up new possibilities in nonlinear optics, where the variation in electric field strength becomes significant for different carrier envelope offset (CEO) phases (Fig. 4). Inside a femtosecond laser oscillator, no coupling mechanism exists between the propagation speeds of the carrier and the pulse envelope. Therefore, the relative delay between the carrier and the envelope of a femtosecond oscillator will exhibit irregular fluctuations unless this jitter is actively suppressed. Different techniques have been proposed to stabilize the CEO frequency in the time domain⁸⁸ and in the frequency domain³⁰. The frequency domain technique is much more sensitive and is generally the technique that is used^{89,90}. We have achieved a long-term CEO stabilization with a residual 10 as timing jitter, which corresponds to 0.025 rad rms CEO phase noise³¹. Applications of CEO nonlinear optics in photoelectron emission³², high harmonic generation³³ and attosecond pulse generation have shown very interesting results³⁴.

Frontier: high average power

Diode-pumped CW lasers could generate hundreds of watts or even kilowatts by the early 1990s; however, the output power of modelocked diode-pumped lasers was still limited to ~100 mW. Only in recent years has this limit increased to 60 W in femtosecond pulses⁶. In terms of pulse energies, 100 mW average output power at a pulse repetition rate of 100 MHz corresponds to 1 nJ of pulse energy. This has been surpassed by three orders of magnitude: the 60 W laser produced nearly 2 μ J pulse energy directly from a SESAM modelocked diode-pumped solid-state laser both in the pico- and femtosecond pulse regime. With external pulse compression, peak powers as high as 12 MW have been generated, as described earlier. In addition, a synchronously pumped fibre-feedback optical parametric oscillator (OPO)⁹¹, a novel type of high-gain synchronously pumped OPO that is particularly insensitive to cavity losses and drifts of the cavity length, produced up to 15 W in 700 fs pulses. OPOs are attractive sources of broadly wavelength-tunable ultrashort pulses, which are required for many applications, for example, RGB display systems.

Initial attempts to obtain higher powers from modelocked lasers had problems, for example, high power scaling was restricted by the requirement to maintain fundamental Gaussian beam quality and to suppress Q-switching instabilities⁹². Many high-power laser heads have large mode areas in the laser gain medium. This often results from the poor beam quality of high-power diode bars. Although for CW lasers this is no problem, it leads to excessive Q-switching instabilities in a passively modelocked laser (Box 3). Thus, more stringent requirements are set for the SESAM design. This is one reason why power scaling typically comes at the expense of longer pulses, because the modulation depth of the SESAM has to be reduced. So far, the best laser head for power scaling of passively modelocked diode-pumped solid-state lasers is based on a thin-disk laser⁹³ and utilizes a disk of, for example, Yb:YAG directly attached to a water-cooled mount. Thermal limitations are minimized by using a disk thickness of 100–250 μ m. Efficient pump absorption in a thin disk is achieved by using multiple (for example, up to 32) passes of the pump radiation. Today, such thin disk lasers are commercially available with more than 1 kW average power. So far Gaussian beam quality has only been demonstrated with up to 100 W of average power, but hundreds of watts should be possible. This combination of SESAM–modelocked thin-disk laser is scalable: the main challenges (beam quality, Q-switching instabilities and SESAM damage) do not become more severe if the power is scaled up. Therefore, the 60 W average power will not remain an upper limit.

Frontier: high pulse repetition rate

Q-switching instabilities limited the highest pulse repetition rates in passively modelocked diode-pumped solid-state lasers to 1 GHz for a long time. For the first time, pulse repetition rates above 10 GHz from

passively modelocked ion-doped solid-state lasers have been generated with Nd:YVO₄ lasers at a centre wavelength around 1 μ m⁹⁴. Shortly afterwards this was increased to 77 GHz⁹⁵ and 160 GHz⁷. Average power has been optimized at a 10 GHz pulse repetition rate to 2.1 W⁷. The peak power was sufficient for efficient nonlinear frequency conversion. For example, a synchronously pumped OPO was demonstrated that produced picosecond pulses that were broadly tunable around 1.55 μ m with up to 100 mW average output power⁹⁶.

Diode-pumped Er:Yb:glass is well suited for telecom applications. Its gain bandwidth covers the entire C-band, it can be pumped with standard 980 nm laser diodes used in erbium-doped fibre amplifiers, and it is robust and inexpensive. However, its small emission cross-section typically prevents high repetition rates without Q-switched modelocking. Using SESAM design optimization this limitation was overcome, and tunable picosecond pulses at 10 GHz¹⁹, 25 GHz²⁵ and 40 GHz²⁰ have been generated.

Actively modelocked fibre lasers can generate pulse repetition rates up to 200 GHz⁹⁷ but only with harmonic modelocking, where multiple pulses circulate in the laser cavity instead of one (Box 1). Good pulse stability is then only achieved by using quite complex means for stabilization. We prefer the simpler approach of fundamental modelocking, that is, with only a single pulse circulating in the laser cavity.

Edge-emitting semiconductor lasers, passively or actively modelocked, can generate repetition rates of more than 1 THz⁹⁸, but have a fairly limited average output power because of the limited mode area. Optically pumped vertical-external cavity surface-emitting lasers (VECSELs) have been passively modelocked with an intracavity saturable absorber⁹⁹. Average power scaling is not limited, and 950 mW of average power in 15 ps pulses with a 6 GHz repetition rate¹⁰⁰ has been generated, with the promise of even higher powers in the multi-gigahertz regime.

The progress in ultrafast compact lasers during the past decade has been simply amazing. The good news is that I strongly believe that this will not be the end of the story. I expect further power scaling to more than 100 W of average power and more compact and cheaper lasers that will support more widespread use for many different applications. □

doi:10.1038/nature01938

- Maria, A. J. D., Stetser, D. A. & Heynau, H. Self mode-locking of lasers with saturable absorbers. *Appl. Phys. Lett.* **8**, 174–176 (1966).
- Keller, U. et al. Solid-state low-loss intracavity saturable absorber for Nd:YLF lasers: an antiresonant semiconductor Fabry–Perot saturable absorber. *Opt. Lett.* **17**, 505–507 (1992).
- Keller, U. et al. Semiconductor saturable absorber mirrors (SESAMs) for femtosecond to nanosecond pulse generation in solid-state lasers. *IEEE J. Sel. Top. Quantum Electron.* **2**, 435–453 (1996).
- Keller, U. in *Nonlinear Optics in Semiconductors* (eds Garmire, E. & Kost, A.) 211–286 (Academic Press, Boston, 1999).
- Spence, D. E., Kean, P. N. & Sibbett, W. 60-fsec pulse generation from a self-mode-locked Ti:sapphire laser. *Opt. Lett.* **16**, 42–44 (1991).
- Innerhofer, E. et al. 60 W average power in 810-fs pulses from a thin-disk Yb:YAG laser. *Opt. Lett.* **28**, 367–369 (2003).
- Kraimer, L. et al. Compact Nd:YVO₄ lasers with pulse repetition rates up to 160 GHz. *IEEE J. Quantum Electron.* **38**, 1331–1338 (2002).
- Shah, J. *Ultrafast spectroscopy of semiconductors and semiconductor nanostructures* (Springer-Verlag, Berlin, 1996).
- Zewail, A. H. Femtochemistry: Recent progress in studies of dynamics and control of reactions and their transition states. *J. Phys. Chem.* **100**, 12701 (1996).
- Zewail, A. H. Femtochemistry: atomic-scale dynamics of chemical bond. *J. Phys. Chem. A* **104**, 5660–5694 (2000).
- Valdmanis, J. A. & Mourou, G. A. Subpicosecond electrooptic sampling: principles and applications. *IEEE J. Quantum Electron.* **22**, 69–78 (1986).
- Weingarten, K. J., Rodwell, M. J. W. & Bloom, D. M. Picosecond optical sampling of GaAs integrated circuits. *IEEE J. Quantum Electron.* **24**, 198–220 (1988).
- Ramaswami, R. & Sivarajan, K. *Optical Networks: A Practical Perspective* (Morgan Kaufmann, 1998).
- Mollenauer, L. F. et al. Demonstration of massive wavelength-division multiplexing over transoceanic distances by use of dispersion-managed solitons. *Opt. Lett.* **25**, 704–706 (2000).
- Miller, D. A. B. Optical interconnects to silicon. *IEEE J. Sel. Top. Quantum Electron.* **6**, 1312–1317 (2000).
- Krishnamoorthy, A. V. & Miller, D. A. B. Scaling optoelectronic-VLSI circuits into the 21st century: a technology roadmap. *IEEE J. Sel. Top. Quantum Electron.* **2**, 55–76 (1996).
- Hatziefremidis, A., Papadopoulos, D. N., Fraser, D. & Avramopoulos, H. Laser sources for polarized electron beams in cw and pulsed accelerators. *Nucl. Instrum. Meth. A* **431**, 46–52 (1999).
- Mollenauer, L. F. & Mamysh, P. V. Massive wavelength-division multiplexing with solitons. *IEEE J. Quantum Electron.* **34**, 2089–2102 (1998).
- Kraimer, L. et al. Tunable picosecond pulse-generating laser with a repetition rate exceeding 10 GHz.

- Electron. Lett.* **38**, 225–227 (2002).
20. Zeller, S. C. *et al.* Passively modelocked 40-GHz Er:Yb:glass laser. *Appl. Phys. B* **76**, 787–788 (2003).
21. Huang, D. *et al.* Optical coherence tomography. *Science* **254**, 1178–1181 (1991).
22. Fujimoto, J. G. Optical coherence tomography. *C. R. Acad. Sci. Paris Serie IV* **2**, 1099–1111 (2001).
23. Boivin, L., Wegmueller, M., Nuss, M. C. & Knox, W. H. 110 Channels x 2.35 Gb/s from a single femtosecond laser. *IEEE Photonics Technology Lett.* **11**, 466–468 (1999).
24. Souza, E. A. D., Nuss, M. C., Knox, W. H. & Miller, D. A. B. Wavelength-division multiplexing with femtosecond pulses. *Opt. Lett.* **20**, 1166–1168 (1995).
25. Spühler, G. J. *et al.* Novel multi-wavelength source with 25-GHz channel spacing tunable over the C-band. *Electron. Lett.* **39**, 778–780 (2003).
26. Holzwarth, R. *et al.* Optical frequency synthesizer for precision spectroscopy. *Phys. Rev. Lett.* **85**, 2264–2267 (2000).
27. Holzwarth, R., Zimmermann, M., Udem, T. & Hänsch, T. W. Optical clockworks and the measurement of laser frequencies with a mode-locked frequency comb. *IEEE J. Quantum Electron.* **37**, 1493–1501 (2001).
28. Stenger, J. *et al.* Phase-coherent frequency measurement of the Ca intercombination line at 657 nm with a Kerr-lens mode-locked femtosecond laser. *Phys. Rev. A* **63**, 021802 (2001).
29. Udem, T., Holzwarth, R. & Hänsch, T. W. Optical frequency metrology. *Nature* **416**, 233–237 (2002).
30. Telle, H. R. *et al.* Carrier-envelope offset phase control: A novel concept for absolute optical frequency measurement and ultrashort pulse generation. *Appl. Phys. B* **69**, 327–332 (1999).
31. Helbing, F. W., Steinmeyer, G., Stenger, J., Telle, H. R. & Keller, U. Carrier-envelope-offset dynamics and stabilization of femtosecond pulses. *Appl. Phys. B* **74**, S35–S42 (2002).
32. Paulus, G. G. *et al.* Absolute-phase phenomena in photoionization with few-cycle laser pulses. *Nature* **414**, 182–184 (2001).
33. Baltuska, A. *et al.* Attosecond control of electronic processes by intense light fields. *Nature* **421**, 611–615 (2003).
34. Drescher, M. *et al.* X-ray pulses approaching the attosecond frontier. *Science* **291**, 1923–1927 (2001).
35. Liu, X., Du, D. & Mourou, G. Laser ablation and micromachining with ultrashort laser pulses. *IEEE J. Quantum Electron.* **33**, 1706–1716 (1997).
36. Nolte, S. *et al.* Ablation of metals by ultrashort laser pulses. *J. Opt. Soc. Am. B* **14**, 2716–2722 (1997).
37. von der Linde, D., Sokolowski-Tinten, K. & Bialkowski, J. Laser-solid interaction in the femtosecond time regime. *Appl. Surf. Sci.* **109/110**, 1–10 (1997).
38. Siders, C. W. *et al.* Detection of nonthermal melting by ultrafast X-ray diffraction. *Science* **286**, 1340–1342 (1999).
39. Rousse, A. *et al.* Non-thermal melting in semiconductors measured at femtosecond resolution. *Nature* **410**, 65–68 (2001).
40. Loesel, F. H., Niemz, M. H., Bille, J. F. & Juhasz, T. Laser-induced optical breakdown on hard and soft tissues and its dependence on the pulse duration: experiment and model. *IEEE J. Quantum Electron.* **32**, 1717–1722 (1996).
41. Hammer, D. X. *et al.* Experimental investigation of ultrashort pulse laser induced breakdown thresholds in aqueous media. *IEEE J. Quantum Electron.* **32**, 670–678 (1996).
42. Juhasz, T. *et al.* Corneal refractive surgery with femtosecond lasers. *IEEE J. Sel. Top. Quantum Electron.* **5**, 902–910 (1999).
43. Loesel, F. H. *et al.* Non-thermal ablation of neural tissue with femtosecond laser pulses. *Appl. Phys. B* **66**, 121–128 (1998).
44. Südmeyer, T. *et al.* Nonlinear femtosecond pulse compression at high average power levels using a large area holey fiber. *Opt. Lett.* (in the press).
45. Ferray, M. *et al.* Multiple-harmonic conversion of 1064 nm radiation in rare gases. *J. Phys. B: At. Mol. Opt. Phys.* **21**, L31–L35 (1988).
46. Lewenstein, M., Balcou, P., Ivanov, M. Y., L’Huillier, A. & Corkum, P. B. Theory of high-harmonic generation by low-frequency laser fields. *Phys. Rev. A* **49**, 2117–2132 (1994).
47. Murnane, M. M., Kapteyn, H. C., Rosen, M. D. & Falcone, R. W. Ultrafast X-Ray pulses from laser-produced plasmas. *Science* **251**, 531–536 (1991).
48. Schmidt, O. *et al.* Time-resolved two-photon photoemission electron microscopy. *Appl. Phys. B* **74**, 223–227 (2002).
49. Bauer, M. *et al.* Direct observation of surface chemistry using ultrafast soft-X-ray pulses. *Phys. Rev. Lett.* **87**, 025501 (2001).
50. Shank, C. V. in *Ultrashort Laser Pulses and Applications* (ed. Kaiser, W.) Chapter 2 (Springer, Heidelberg, 1988).
51. Diels, J.-C. in *Dye lasers principles: with applications* (eds Duarte, F. J. and Hillman, L. W.) 41–132 (Academic Press, Boston, 1990).
52. Shank, C. V. & Ippen, E. P. Subpicosecond kilowatt pulses from a modelocked cw dye laser. *Appl. Phys. Lett.* **24**, 373–375 (1974).
53. Fork, R. L., Cruz, C. H. B., Becker, P. C. & Shank, C. V. Compression of optical pulses to six femtoseconds by using cubic phase compensation. *Opt. Lett.* **12**, 483–485 (1987).
54. Moulton, P. F. Spectroscopic and laser characteristics of Ti:Al₂O₃. *J. Opt. Soc. Am. B* **3**, 125–132 (1986).
55. Ishida, Y., Sarukura, N. & Nakano, H. in *Ultrafast Phenomena PD11-11* (OSA, California, 1990).
56. Spence, D. E., Kean, P. N. & Sibbett, W. in *Conference on lasers and electro-optics (CLEO) CPDP10* (OSA/IEEE LEOS, Anaheim, California, 1990).
57. Keller, U., ‘tHooft, G. W., Knox, W. H. & Cunningham, J. E. Femtosecond pulses from a continuously self-starting passively mode-locked Ti:Sapphire laser. *Opt. Lett.* **16**, 1022–1024 (1991).
58. Salin, F., Squier, J. & Piché, M. Modelocking of Ti:Sapphire lasers and self-focusing: a Gaussian approximation. *Opt. Lett.* **16**, 1674–1676 (1991).
59. Negus, D. K., Spinelli, L., Goldblatt, N. & Feugnet, G. in *Advanced Solid-State Lasers* (eds Dubé, G. & Chase, L.) 120–124 (Optical Society of America, Washington D.C., 1991).
60. Keller, U., Knox, W. H. & ‘tHooft, G. W. Ultrafast solid-state modelocked lasers using resonant nonlinearities. *IEEE J. Quantum Electron.* **28**, 2123–2133 (1992).
61. Valdmantis, J. A. & Fork, R. L. Design considerations for a femtosecond pulse laser balancing self phase modulation, group velocity dispersion, saturable absorption, and saturable gain. *IEEE J. Quantum Electron.* **22**, 112–118 (1986).
62. Sutter, D. H. *et al.* Semiconductor saturable-absorber mirror-assisted Kerr-lens mode-locked Ti:Sapphire laser producing pulses in the two-cycle regime. *Opt. Lett.* **24**, 631–633 (1999).
63. Ell, R. *et al.* Generation of 5-fs pulses and octave-spanning spectra directly from a Ti:Sapphire laser. *Opt. Lett.* **26**, 373–375 (2001).
64. Cerullo, G., De Silvestri, S., Magni, V. & Pallaro, L. Resonators for Kerr-lens mode-locked femtosecond Ti:Sapphire lasers. *Opt. Lett.* **19**, 807–809 (1994).
65. Cerullo, G., De Silvestri, S. & Magni, V. Self-starting Kerr lens mode-locking of a Ti:Sapphire laser. *Opt. Lett.* **19**, 1040–1042 (1994).
66. Gibson, A. F., Kimmitt, M. F. & Norris, B. Generation of bandwidth-limited pulses from a TEA CO₂ laser using p-type germanium. *Appl. Phys. Lett.* **24**, 306–307 (1974).
67. Ippen, E. P., Eichenberger, D. J. & Dixon, R. W. Picosecond pulse generation by passive modelocking of diode lasers. *Appl. Phys. Lett.* **37**, 267–269 (1980).
68. Islam, M. N. *et al.* Color center lasers passively mode locked by quantum wells. *IEEE J. Quantum Electron.* **25**, 2454–2463 (1989).
69. Hönninger, C., Paschotta, R., Morier-Genou, F., Moser, M. & Keller, U. Q-switching stability limits of continuous-wave passive mode locking. *J. Opt. Soc. Am. B* **16**, 46–56 (1999).
70. Paschotta, R. & Keller, U. Passive mode locking with slow saturable absorbers. *Appl. Phys. B* **73**, 653–662 (2001).
71. Kärtner, F. X., Jung, I. D. & Keller, U. Soliton Modelocking with Saturable Absorbers. *IEEE J. Sel. Top. Quantum Electron.* **2**, 540–556 (1996).
72. Ell, R. *et al.* Generation of 5-fs pulses and octave-spanning spectra directly from a Ti:Sapphire laser. *Opt. Lett.* **26**, 373–375 (2001).
73. Nisoli, M. *et al.* Compression of high energy laser pulses below 5 fs. *Optics Lett.* **22**, 522–524 (1997).
74. Shirakawa, A., Sakane, I., Takasaka, M. & Kobayashi, T. Sub-5-fs visible pulse generation by pulse-front-matched noncollinear optical parametric amplification. *Appl. Phys. Lett.* **74**, 2268–2270 (1999).
75. Gallmann, L. *et al.* Characterization of sub-6-fs optical pulses with spectral phase interferometry for direct electric-field reconstruction. *Opt. Lett.* **24**, 1314–1316 (1999).
76. Schenkel, B. *et al.* Generation of 3.8-fs pulses from adaptive compression of a cascaded hollow fiber supercontinuum. *Opt. Lett.* (in the press).
77. Gallmann, L., Sutter, D. H., Matuschek, N., Steinmeyer, G. & Keller, U. Techniques for the characterization of sub-10-fs optical pulses: a comparison. *Appl. Phys. B* **70**, S67–S75 (2000).
78. Fork, R. L., Martinez, O. E. & Gordon, J. P. Negative dispersion using pairs of prisms. *Opt. Lett.* **9**, 150–152 (1984).
79. Asaki, M. T. *et al.* Generation of 11-fs pulses from a self-mode-locked Ti:Sapphire laser. *Opt. Lett.* **18**, 977–979 (1993).
80. Curley, P. F. *et al.* Operation of a femtosecond Ti:Sapphire solitary laser in the vicinity of zero group-delay dispersion. *Opt. Lett.* **18**, 54–56 (1993).
81. Zhou, J. *et al.* Pulse evolution in a broad-bandwidth Ti:Sapphire laser. *Opt. Lett.* **19**, 1149–1151 (1994).
82. Szpöcs, R., Ferencz, K., Spielmann, C. & Krausz, F. Chirped multilayer coatings for broadband dispersion control in femtosecond lasers. *Opt. Lett.* **19**, 201–203 (1994).
83. Stingl, A., Lenzner, M., Ch. Spielmann, Krausz, F. & Szpöcs, R. Sub-10-fs mirror-dispersion-controlled Ti:Sapphire laser. *Opt. Lett.* **20**, 602–604 (1995).
84. Kärtner, F. X. *et al.* Design and fabrication of double-chirped mirrors. *Opt. Lett.* **22**, 831–833 (1997).
85. Matuschek, N., Kärtner, F. X. & Keller, U. Analytical design of double-chirped mirrors with custom-tailored dispersion characteristics. *IEEE J. Quantum Electron.* **35**, 129–137 (1999).
86. Matuschek, N., Gallmann, L., Sutter, D. H., Steinmeyer, G. & Keller, U. Back-side coated chirped mirror with ultra-smooth broadband dispersion characteristics. *Appl. Phys. B* **71**, 509–522 (2000).
87. Steinmeyer, G., Sutter, D. H., Gallmann, L., Matuschek, N. & Keller, U. Frontiers in ultrashort pulse generation: pushing the limits in linear and nonlinear optics. *Science* **286**, 1507–1512 (1999).
88. Xu, L. *et al.* Route to phase control of ultrashort light pulses. *Opt. Lett.* **21**, 2008–2010 (1996).
89. Jones, D. J. *et al.* Carrier-envelope phase control of femtosecond mode-locked lasers and direct optical frequency synthesis. *Science* **288**, 635–639 (2000).
90. Apolonski, A. *et al.* Controlling the phase evolution of few-cycle light pulses. *Phys. Rev. Lett.* **85**, 740 (2000).
91. Südmeyer, T. *et al.* Femtosecond fiber-feedback OPO. *Opt. Lett.* **26**, 304–306 (2001).
92. Paschotta, R. *et al.* Diode-pumped passively mode-locked lasers with high average power. *Appl. Phys. B* **70**, S25–S31 (2000).
93. Giesen, A. *et al.* Scalable concept for diode-pumped high-power solid-state lasers. *Appl. Phys. B* **58**, 363–372 (1994).
94. Krainer, L. *et al.* Passively mode-locked Nd:YVO₄ laser with up to 13 GHz repetition rate. *Appl. Phys. B* **69**, 245–247 (1999).
95. Krainer, L., Paschotta, R., Moser, M. & Keller, U. 77 GHz soliton modelocked Nd:YVO₄ laser. *Electron. Lett.* **36**, 1846–1848 (2000).
96. Lecomte, S. *et al.* Optical parametric oscillator with 10 GHz repetition rate and 100 mW average output power in the 1.5-μm spectral region. *Opt. Lett.* **27**, 1714–1717 (2002).
97. Yoshida, E. & Nakazawa, M. 80–200GHz erbium doped fibre laser using a rational harmonic mode-locking technique. *Electron. Lett.* **32**, 1370–1372 (1996).
98. Arahira, S., Matsui, Y. & Ogawa, Y. Mode-locking at very high repetition rates more than terahertz in passively mode-locked distributed-Bragg-reflector laser diodes. *IEEE J. Quantum Electron.* **32**, 1211–1224 (1996).
99. Hoogland, S. *et al.* Passively mode-locked diode-pumped surface-emitting semiconductor laser. *IEEE Photon. Technol. Lett.* **12**, 1135–1138 (2000).
100. Häring, R. *et al.* High-power passively mode-locked semiconductor lasers. *IEEE J. Quantum Electron.* **38**, 1268–1275 (2002).

Acknowledgements I would like to acknowledge many important contributions from my graduate students, post-docs and collaborators. I would like to mention Rüdiger Paschotta, Günter Steinmeyer, Franz Kärtner, Markus Haiml who have been senior researchers and project leaders in my group at different times during the last 10 years and Kurt Weingarten, Gabriel Spühler and Lukas Krainer from GigaTera Inc. on the most recent work on high pulse repetition rate lasers. This work has been supported by ETH, the Swiss National Science Foundation (SNF) and the Swiss Technology and Innovation Foundation (KTI).

Optical microcavities

Kerry J. Vahala

California Institute of Technology, Mail Stop 128-95, Pasadena, California 91125, USA (e-mail: vahala@caltech.edu)

Optical microcavities confine light to small volumes by resonant recirculation. Devices based on optical microcavities are already indispensable for a wide range of applications and studies. For example, microcavities made of active III–V semiconductor materials control laser emission spectra to enable long-distance transmission of data over optical fibres; they also ensure narrow spot-size laser read/write beams in CD and DVD players. In quantum optical devices, microcavities can coax atoms or quantum dots to emit spontaneous photons in a desired direction or can provide an environment where dissipative mechanisms such as spontaneous emission are overcome so that quantum entanglement of radiation and matter is possible. Applications of these remarkable devices are as diverse as their geometrical and resonant properties.

Like its acoustic analogue the tuning fork, the optical microcavity (or microresonator) has a size-dependent resonant frequency spectrum. Microscale volume ensures that resonant frequencies are more sparsely distributed throughout this spectrum than they are in a corresponding ‘macroscale’ resonator. An ideal cavity would confine light indefinitely (that is, without loss) and would have resonant frequencies at precise values. Deviation from this ideal condition is described by the cavity Q factor (which is proportional to the confinement time in units of the optical period). Q factor and microcavity volume (V) figure prominently in applications of these devices, and a summary of values typical for the devices discussed in this review is given in Table 1. In addition, representative examples of the three methods of confinement employed in microcavities are provided in Figs 1–3 (refs 1–7).

In this review, I consider four applications of optical microcavities: strong-coupling cavity quantum electrodynamics (QED), enhancement and suppression of spontaneous emission, novel sources, and dynamic filters in optical communication. These areas are just four of several possible, and many topics, such as soliton effects^{8,9}, chaos¹⁰ and effects in quantum-well microcavities¹¹, will not be reviewed because of space limitations. Also, I will not review microcavity types in commercial semiconductor lasers because extensive texts and treatises on this subject already exist^{12,13}. Even for the four applications discussed, there are by necessity omissions. Cavity QED, for example, is a vast topic, and selections have been made on the basis of their importance and for the interesting design limits they illustrate. I will provide a brief introduction to each area, then describe a few representative applications, their microcavity requirements, and the state-of-the-art for these devices, before outlining the challenges for the future.

Strong-coupling cavity QED

An electron transitions within an atom from an excited state to a ground state, emitting a photon into the continuum of radiation modes¹⁴. This irreversible behaviour is an example of weak coupling. If the same atom is inserted into a microcavity and if the strong-coupling conditions^{15–17} described in Box 1 are satisfied, then the atom can interact coherently with a cavity mode for a meaningful time. This can take place even when the mode is initially in its vacuum state (ground state or state of lowest energy). Use of a weak optical probe reveals that the cavity’s transmission spectrum is split by the presence of the atom into two distinct peaks, which correspond to eigen frequencies of the quantum entangled atom-

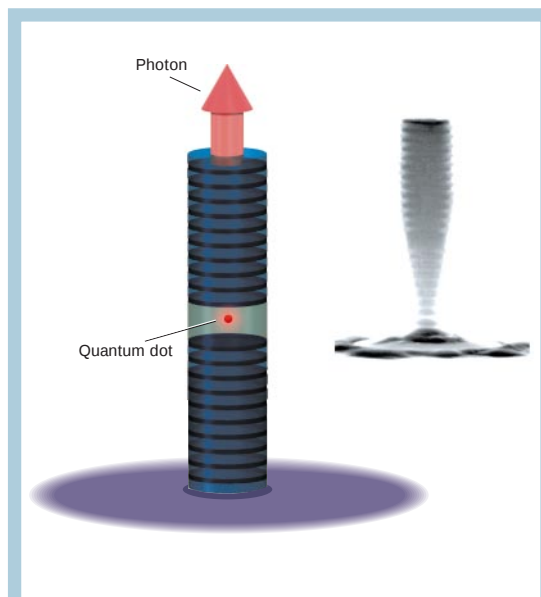


Figure 1 Micropost (or micropillar) cavities^{1,2} have played a major role in recent applications of the Purcell effect to triggered, single-photon sources. They offer small cavity volume and relatively high Q , have an emission pattern that is well suited for coupling and manipulation of emitted photons⁵⁶ (for example, with optical fibres) and can incorporate quasi-atomic, quantum dots as emitters. In the rendering, Bragg mirrors³ at the output (upper stack near arrow) and below provide one dimension of cavity confinement, whereas air-dielectric guiding provides lateral (in-the-plane) confinement. A single quantum dot is shown spontaneously emitting a photon that is subsequently directed via the Purcell effect through the cavity top. The inset shows a scanning electron micrograph of such a micropost cavity used in recent triggered single-photon source experiments. Inset micrograph courtesy of Y. Yamamoto⁷⁶ (Stanford University, CA).

cavity states (states that are not factorable into cavity and atom components)¹⁸. If the probe frequency is maintained at the cavity’s original resonant frequency, then the entry of a single atom into the cavity can block transmission¹⁹ (that is, the probe is reflected). In state-of-the-art systems, the resulting extreme sensitivity of transmission to the atom’s position within the cavity is used to ascertain atomic centre-of-mass motion^{20–22}, such as the orbital motion of ultracold atoms entrained by their interaction with the cavity mode^{21,22} (Fig. 4). In addition, such a cavity containing cold atoms exhibits extreme dispersive properties, which have

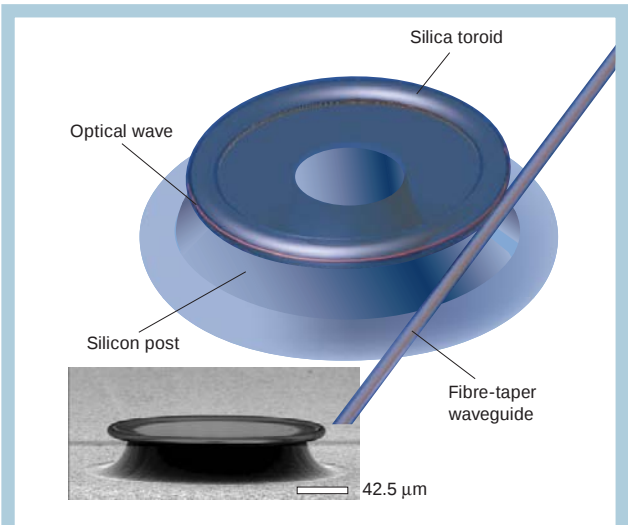


Figure 2 Rendering of an ultrahigh- Q microtoroid resonator⁶. An optical wave, shown in red, is coupled from a fibre-taper waveguide and subsequently guided within and along the periphery of the microtoroid in a whispering gallery mode, which is named after its acoustic equivalent⁵. Whispering gallery microcavities can be found in several geometries including spheres (see Fig. 5), disks (see Fig. 6) and rings (inset to Fig. 6). Inset: A scanning electron micrograph of a microtoroid resonator consisting of a thin silica layer upon a silicon post and substrate. The device has a diameter of 120 μm and exhibits a Q factor in excess of 100 million. The smooth exterior toroid surface is the result of the toroid going through a molten state during its fabrication. Inset micrograph courtesy of D. Armani⁶.

recently been observed²³. These measurements, as well as applications of cavity QED to quantum information studies, have recently been reviewed elsewhere^{24,25}.

Efforts to increase strong-coupling effects (as measured by reductions in saturation photon number, n_0 , and critical atom number, N_0 (see Box 1)) have provided much impetus for research into ultrahigh- Q , small-volume resonant cavities^{17,26–30}. Both n_0 and N_0 have been lowered from near unity levels in the earliest demonstrations of single-atom vacuum Rabi splitting at optical frequencies to recent levels of $n_0 = 2.82 \times 10^{-4}$ and $N_0 = 6.1 \times 10^{-3}$ (refs 21,31). Microcavities in these experiments feature Fabry–Perot-style resonators (an optical resonator in which feedback is accomplished using two mirrors) with ultrahigh reflectance mirror technology²⁶. A cavity finesse, an alter-

native measure of cavity perfection that does not include propagation effects within the cavity as does the Q factor, of 1.9×10^5 (ref. 26) has been obtained using these mirrors. Optimization of n_0 and N_0 , however, involves joint optimization of the mode volume and finesse (or Q). So, for example, the results cited above used a resonator with mode volume $V = 1.69 \times 10^3 \mu\text{m}^3$ and finesse of 4.8×10^5 . A detailed review of the technological limits imposed by mirror technology in optimizing Fabry–Perot microcavities for strong-coupling studies has recently been performed³².

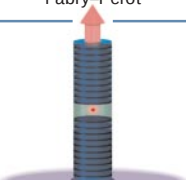
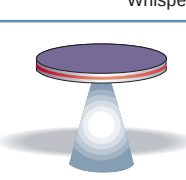
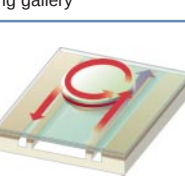
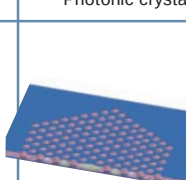
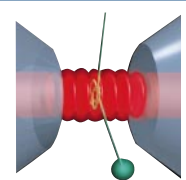
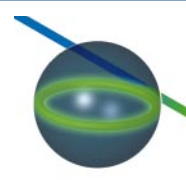
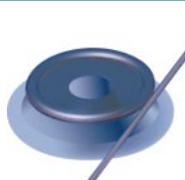
In addition to ultrahigh-finesse Fabry–Perot microcavities, the whispering gallery modes of silica and quartz microspheres have received considerable attention^{27–29,33,34}. Whispering gallery resonators are typically dielectric spherical structures in which waves are confined by ‘continuous total internal reflection’. Silica microspheres, which are robust ultrahigh- Q microresonators, were first studied by Braginsky and Ilchenko³⁵. Spheres feature an atomic-like mode spectrum in which high ℓ number (principal angular index or optical mode) and low radial number modes execute orbits near the sphere’s surface³³ (Fig. 5). Excellent surface finish is crucial for maximizing Q , and the formation of spheres through surface tension (that is, as a molten droplet) provides a near atomically smooth surface (with only a few nanometres or less of surface roughness^{28,29}). The bulk optical loss from silica is also exceptionally low and record Q factors^{28,29} of 8×10^9 (and finesse²⁹ of 2.3×10^6) have been obtained. For these measurements, dependence of Q on sphere diameter is consistent with Q being limited by losses of surface roughness²⁹. Also, a time dependency for the measured Q was observed and is believed to result from water adsorption and formation of OH groups at the sphere’s surface^{28,34}. For diameters below 20 μm in silica spheres, radiation leakage becomes a significant factor in determining Q ³¹. The lowest order radial modes (in terms of nodes) with $m = \ell^{31}$ are minimal volume, equatorial ring orbits (see Fig. 5) and are best suited for cavity QED. Experimental work has demonstrated strong coupling in this system³⁴, and recent modelling³¹ shows that substantial improvements in strong coupling are possible using spheres with reduced diameters.

Microcavities based on photonic crystals (Fig. 3) can provide extremely small mode volumes⁷, and donor-mode cavity geometries (in which a small additional hole is drilled within the design of Fig. 3) have been modelled with a neutral atom suspended within the hole³⁰. Strong coupling is theoretically feasible; however, at present, Q values in fabricated structures are well below the theoretical optima.

For the purposes of optical probing/output-coupling, Fabry–Perot cavities enable direct ‘endfire’ coupling along the cavity axis. Whispering gallery modes, however, must be phase matched³⁶,

Table 1

The microcavities are organized by column according to the confinement method used and by row according to high Q and ultrahigh Q . Small mode volume and ultrasmall mode volume are other possible classifications that are somewhat complementary to this scheme. Representative, measured Q s and V s are given and have been taken from the following cited references. Upper row: micropost³⁸, microdisk³², semiconductor¹⁰³, polymer¹⁰⁴ add/drop filter, photonic crystal cavity⁶². Lower row: Fabry–Perot bulk optical cavity^{21,31}, microsphere²⁹, microtoroid⁶. n is the material refractive index, and, V , if not indicated, was not available. Microsphere volume V was inferred using the diameter noted in the cited reference and finesse (F) is given for the ultrahigh- Q Fabry–Perot as opposed to Q . Two Q values are cited for the add/drop filter: one for a polymer design, Q_{poly} , and the second for a III–V semiconductor design, $Q_{\text{III-V}}$.

	Fabry–Perot	Whispering gallery	Photonic crystal
High Q	 $Q: 2,000$ $V: 5 (\lambda/n)^3$	 $Q: 12,000$ $V: 6 (\lambda/n)^3$  $Q_{\text{III-V}}: 7,000$ $Q_{\text{poly}}: 1.3 \times 10^5$	 $Q: 13,000$ $V: 1.2 (\lambda/n)^3$
Ultrahigh Q	 $F: 4.8 \times 10^5$ $V: 1,690 \mu\text{m}^3$	 $Q: 8 \times 10^9$ $V: 3,000 \mu\text{m}^3$  $Q: 10^8$	

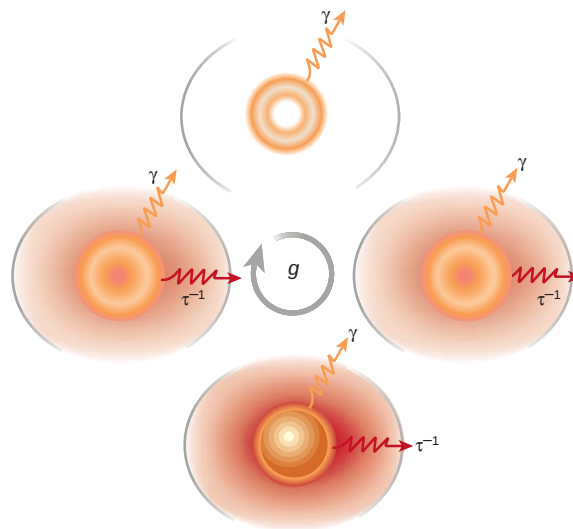
Box 1 Strong coupling

An atom, initially in the excited state of a dipole transition, enters a lossless microcavity of volume V . A single mode of the cavity, in its ground state, is resonant with the transition. The atom and 'vacuum field' couple, which results in a quantum of energy shifting back and forth between the atom and the mode at the vacuum Rabi frequency. The interaction strength of atom and cavity mode is linear in the field and hence smaller cavity volumes concentrate the vacuum field of the mode, producing larger Rabi frequencies. This fundamental dynamic of the atom–field system is reversible as long as the system is isolated. In reality the cavity will have a finite photon lifetime (finite Q) that will limit, perhaps even prevent, Rabi oscillations by allowing the energy to leak irreversibly into the continuum. Likewise, the atomic transition will couple to continuum radiation modes and thereby experience spontaneous decay of its population as well as polarization dephasing. Strongly coupled systems are those in which the Rabi dynamic can exist, even if only briefly, despite the reality of dissipation. Strong coupling occurs when the atom–field coupling strength, g (which is half the Rabi frequency), is faster than any underlying dissipative rate and larger than $1/T$ where T is the interaction time¹⁷. Under conditions of strong coupling, weak optical probing near the microcavity resonant frequency reveals two spectral transmission peaks (where only one existed before) giving the energies of new eigen states, which are now entangled states of the atom and cavity field. Given modal and atomic dissipation, the degree to which the atom and cavity mode are strongly coupled can be quantified by defining a saturation photon number (n_o) and critical atom number (N_o). These quantities are, by necessity, functions of parameters describing both the reversible dynamic and the various decay rates¹⁷,

$$n_o \propto \frac{\gamma_{\parallel}\gamma_{\perp}}{g^2} \propto V \quad N_o \propto \frac{\gamma_{\perp}}{\omega g^2} \propto \frac{V}{Q}$$

where $(\gamma_{\parallel}, \gamma_{\perp})$ are the atomic dissipation rates (population relaxation and atomic dephasing, respectively) and τ is the cavity lifetime. In

1992, unity n_o and N_o were demonstrated at optical frequencies¹⁸. Recently, state-of-the-art, strongly coupled systems have produced n_o and N_o values that are much less than unity, reflecting a fundamental Rabi dynamic that exists over many cycles. For reviews of strong coupling and its application to cavity QED see refs 17,25,46,120.



Box 1 Figure Vacuum Rabi oscillation. An excited atom is introduced into the cavity (top) and undergoes vacuum Rabi oscillation mediated by the atom–field coupling strength, g , resulting in one quantum being added to the mode (shown in red) (bottom). Dissipation mechanisms are also illustrated. γ is the atomic damping rate and τ is the cavity lifetime.

which is typically achieved using total internal reflection from the back face of a prism. Other coupling methods have also been demonstrated^{37–40}. For many applications of cavity QED, such as in quantum information studies, parasitic loss in coupling to and from microresonators will figure prominently. A fibre-optic cable is considered a likely medium over which to transport quantum information⁴¹. In this regard, fibre-optic tapers (see Fig. 5) provide ultralow loss, direct coupling to ultrahigh- Q spheres⁴² and have been proposed as a means to couple quantum states to or from a resonator onto a fibre^{42,43}. Also, the recent demonstration of a fibre-taper-coupled ultrahigh- Q microtoroid-on-a-chip⁶ (see Fig. 2) enables integration of wafer-based functions with ultralow-loss fibre-coupled quantum devices.

Enhancement and suppression of spontaneous emission

Weak coupling results when dissipation overwhelms the fundamental Rabi dynamic. The control of spontaneous emission through the Purcell effect⁴⁴ in this regime has become an important application of microcavities⁴⁵. Because all cavities exhibit loss at some level, cavity modes are, more rigorously, quasi modes, and, in the strictest sense, the spectral–lineshape function of a ‘discrete’ mode is proportional to a continuum density of modes (DM) function (ref. 4, chapter 1, and references therein). Applications of the Purcell effect make use of this duality by, on the one hand, using local enhancement of the continuum DM function to influence spontaneous emission, while, on the other, using the quasi mode of the resonator as the target of the emission process (see Box 2 for further discussion of this and the complementary effect, spontaneous-emission suppression). Thus, atomic decay rates are not only speeded-up but are also made direc-

tional. Reviews of this mechanism both at optical and at microwave frequencies appear in ref. 4 (chapter 2) and ref. 46.

The ability to create InAs quantum dots that have excellent luminescent properties⁴⁷ within III–V semiconductors was a turning point in the recent history of the Purcell effect. As noted by Gerard¹, these structures can act as a local probe of the field within a III–V microcavity and can also efficiently capture and then confine electrons and holes, making them less susceptible to the semiconductor surface effects that occur when cavity dimension decreases. Quantum dots are also well suited to the study of the Purcell effect in semiconductors, because, unlike bulk or quantum-well media, individual quantum dots exhibit a relatively narrow spectral lineshape that fits within a high- Q microcavity mode^{48,49}.

Using quantum-dot-loaded micropost^{1,3} (or micropillar) cavities, such as the one shown in Fig. 1, with Q s of 2,000 (1 μm diameter posts), Gerard and coworkers showed a five-fold Purcell enhancement^{48,50}. Spontaneous emission suppression was later verified in similar structures using a metallic sidewall coating to exclude transverse continuum modes⁵¹. Purcell enhancement of emission from a single quantum dot within a micropost was later demonstrated². In addition to microposts, microdisk cavities incorporating quantum dots support modes with Q s in the range of 10,000–17,000^{52,53}, and Purcell enhancement has been measured⁵⁴ up to 15-fold⁵⁵ in these structures. Measured Purcell factors are influenced by radial spatial averaging, which occurs over an ensemble of resonant emitters^{48,56}, and estimated Purcell factors based on Q measurement and cavity volume estimates are typically much higher. Using Q measurement alone, a Purcell factor of 32 for microposts⁵⁰ and a factor of

Box 2 Purcell effect

A two level system will decay spontaneously by interaction with a vacuum continuum at a rate proportional to the spectral density of modes per volume evaluated at the transition frequency. Within a cavity, the density of modes is modified and large swings in its amplitude can occur. From the viewpoint of cavity modes (which in the presence of dissipation must be viewed as quasi modes (ref. 4, chapter 1)), the maximal density of modes occurs at the quasi-mode resonant frequencies and can greatly exceed the corresponding free-space density. Historically, Purcell⁴⁴ arrived at this conclusion by noting that a single (quasi) mode occupies a spectral bandwidth ν/Q within a cavity of volume V . Normalizing a resulting cavity-enhanced mode density per unit volume to the mode density of free space yields the 'Purcell' spontaneous emission enhancement factor^{44,46}.

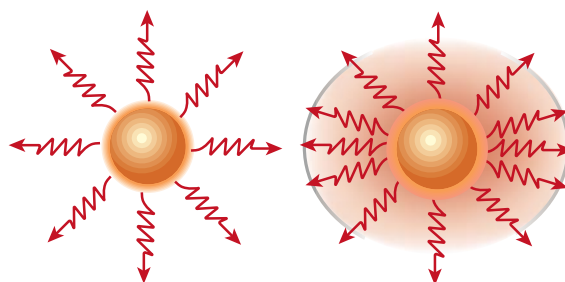
$$P = \frac{3}{4\pi^2} \left(\frac{\lambda}{n} \right)^3 \frac{Q}{V}$$

where refractive index, n , is a modern addition to this expression to account for emission within dielectrics⁵⁵. An atom whose transition falls within the mode linewidth will experience an enhancement to its spontaneous decay rate given by the Purcell factor. More significantly, because the enhancement comes about from coupling to only those continuum modes that make up the corresponding quasi-mode of the resonator, the spontaneous emission is directed to this quasi mode⁸⁸ and has great utility with regard to coupling spontaneous power.

In spectral locations that are intermediate to modal resonance frequencies, the density of the modes can fall well below the density in free space. With proper cavity design and for operation at these off-resonance frequencies, spontaneous decay can be suppressed^{4,46,58}. Rigorous developments of Purcell's physical model have been

presented on the basis of calculation of the continuum mode density (ref. 4, chapters 1 and 2 and references therein).

Design of microcavities for observation of the Purcell effect must take account of the corresponding atomic (or atom-like) transition characteristics. Use of a small microcavity volume is important because enhancements driven by manipulation of Q alone are limited by the spectral width of the transition. Likewise, all other things being equal, narrow, atomic transitions can be Purcell-enhanced more as a higher Q becomes possible. It is for this reason that individual quantum dots, with their relatively narrow transition widths (compared with bulk semiconductors), are playing a significant role in this field⁴⁹ (see Fig. 1).



Box 2 Figure Purcell enhancement of spontaneous emission. Weak coupling to a cavity mode will enhance the spontaneous rate of emission by increasing the local density of modes (right) compared with their density in free space (left).

190 for 2 μm diameter microdisks⁵² have been inferred. Q factors as high as 10,000 with corresponding mode volumes of $1.6 (\lambda/n)^3$ have been predicted to exist in optimized micropost cavities⁵⁷. A post diameter of 0.5 μm with an improved Q factor of 4,800 yields a Purcell factor of 147^{56,57}.

Since Yablonovitch first proposed using a photonic crystal for spontaneous emission suppression⁵⁸, much attention has been directed to photonic bandgap microcavities^{59,60}. As noted earlier, photonic-crystal defect microcavities can provide extremely small mode volumes⁷, and large theoretical Q values have been predicted for certain designs^{30,61}. Recently, a Q of 4,000 for an H2 (seven holes removed to form the hexagon) defect cavity⁵⁹ and a Q of 13,000⁶² for a donor-mode cavity (calculated mode volume of $1.2 (\lambda/n)^3$) were reported. Purcell enhancement has also been studied in this system^{60,63}.

Controlling the emission of single photons has been a priority for quantum encryption systems⁶⁴. Single-photon sources, which are required in these systems, are a recent application of the Purcell effect in quantum-dot microcavities (see Fig. 1). Quantum dots are quasi-atomic systems and hence share many properties with atoms. For example, emission from a single atom or molecule and from quantum dots⁶⁷ exhibits non-classical photon anti-bunching behaviour, because, upon emission, an interval must pass in order for the atom to be re-excited and to emit a photon^{65,66}. This behaviour in quantum dots has been adapted to generate triggered single photons⁶⁸⁻⁷¹. Leading up to this application, triggered single-photon emission using photo-pumped, single-molecule systems was demonstrated^{72,73}. However, quantum-dot single-photon sources, which are compact and potentially electrically pumped, are very appealing for many of the same reasons that semiconductor lasers are so compelling in communications. Unlike lasers, however, the useful emission in these new quantum sources is a spontaneous photon; therefore, stimulated emission cannot be relied upon to direct

power into a single cavity mode (a necessity for efficient coupling to optical fibres). Instead, the Purcell effect is applied to improve coupling⁵⁵. Both microdisk⁶⁸ and micropost-based devices⁷⁴⁻⁷⁶ have been demonstrated. Significantly, a single photon source that is an electrical-pumped single quantum dot has recently been demonstrated⁷⁷. An

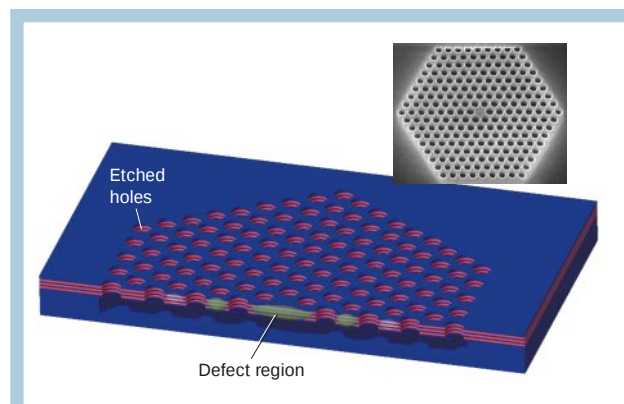


Figure 3 Cross-sectional illustration of a photonic crystal defect microcavity laser. The microcavity is formed by dry etching a hexagonal array of holes and subsequent selective etch of an interior region, creating a thin membrane. One hole is left unetched creating a 'defect' in the array and therefore a defect mode in the optical spectrum. The mode (illustrated in green) is confined to the interior of the array by Bragg reflection in the plane and conventional waveguiding in the vertical direction. Also, shown in pink are quantum wells that upon photo pumping provide the amplification necessary for laser oscillation⁷. Inset: Scanning electron micrograph of a photonic crystal defect microcavity laser. Micrograph is courtesy of O. Painter and A. Scherer (Caltech, CA).

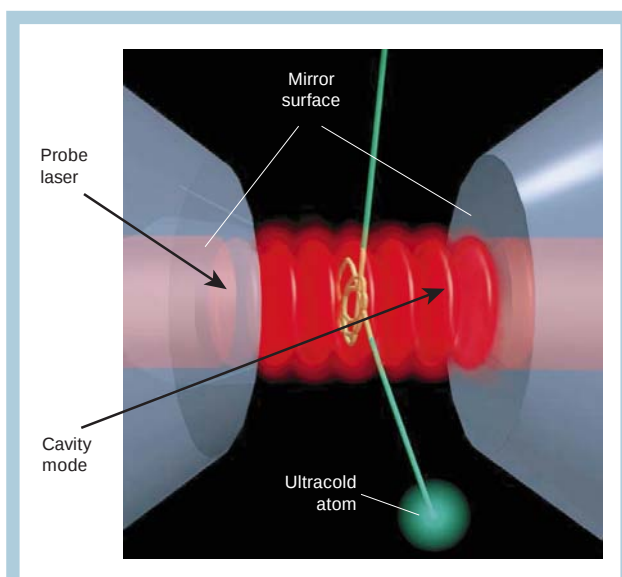


Figure 4 If the coupling energy $\hbar g$ in a strongly coupled system exceeds the thermal energy of the atom, then the atomic centre of mass motion will be altered by interaction with the vacuum cavity mode. In recent experiments^{21,22}, ultracold atoms have been produced to satisfy this condition, resulting in atomic motion that is entrained for substantial periods of time by cavity QED coupling. In this figure, an ultracold atom is entrained in an orbital motion before escaping. Because the coupling energy depends on the amplitude of the vacuum cavity field near the atom, optical transmission probing of the cavity during the atomic entrainment acts as an ultrasensitive measure of atomic location. Figure used with permission of H. J. Kimble (Caltech, CA).

important next step will be to combine this feature with a microcavity equipped for significant Purcell enhancement.

Novel sources

The pursuit of efficient and compact laser sources that offer enhanced functionality or that provide insight into microcavity physics has inspired a large body of microcavity research. Lasing has been demonstrated in droplets^{78,79}, silica^{38,80} and polystyrene spheres⁸¹, semiconductor microdisks^{52,82,83}, micropillars³ (vertical cavities⁸⁴) and photonic crystal cavities⁷. Small cavity volumes and high Q have allowed the production of submicrowatt optical pump thresholds⁸⁵ in microspheres and microamp-scale current thresholds in semiconductor lasers^{86,87}. With the advent of multi-wavelength communications systems¹², tunable and compact sources have taken on an increased importance. In addition, interest in the Purcell effect and more efficient lasers has focused attention on threshold control⁸⁸ and also on the concept of threshold⁸⁹. Lasers⁹⁰ that operate like micro-masers^{15,46} have also been studied.

A development of practical importance is the use of lateral oxidation in vertical cavity lasers⁸⁴. These lasers have a lateral oxide aperture that is normal to the cavity axis, which creates lateral mode confinement and concentrates pumping current at the optical gain region, thereby making the device very efficient^{86,87}. Cavity enhancement effects have also been observed in versions of these devices containing quantum dots⁹¹.

Sources that use a nonlinearly stimulated process to achieve laser action represent another class of device. Resonant recirculation of weak input signals within ultrahigh- Q , small-mode-volume resonators will produce enormous modal field intensities and thereby lower the threshold for nonlinear phenomena³⁵. For a given coupled, input power, P_{in} , the circulating intensity within the resonator is given by $I = P_{in} / (\lambda / 2\pi n) (Q/V)$ where n is the group index. For a cavity Q of 100 million and a mode volume of $500 \mu\text{m}^3$ (both obtainable in spheres roughly $40 \mu\text{m}$ in diameter^{31,43}) the circulating intensity

exceeds 1 GWatt/cm^2 with less than 1 mW of coupled input power. Observation of stimulated Raman scattering^{92,93}, multi-order Stokes emission⁹⁴, stimulated Brillouin scattering⁹⁵ and many other nonlinear effects were first studied in microdroplets by Chang⁹² and by Campillo (ref. 4, chapter 5, and references therein). The Kerr effect has also been observed by Treussart *et al.* in ultrahigh- Q silica microspheres⁹⁶ at microwatt input power levels. More recently, efficient solid-state Raman laser sources using fibre-coupled⁴² ultrahigh- Q microspheres have been demonstrated, and they produced record-low-threshold pump powers of $65 \mu\text{W}$ (ref. 43) (see Fig. 5). Raman sources can be used to extend the wavelength range of conventional lasers into difficult-to-access bands.

Dynamic filters in optical communications

During the past decade, wavelength division multiplexed (WDM) lightwave systems have been deployed in long-distance transmission to take better advantage of the vast bandwidth available in an optical fibre⁹⁷. Beyond providing additional bandwidth, wavelength, used as a dynamic parameter, can enhance network performance⁹⁸. With attention turned towards future system needs, there has been interest in devices that enable new filtering and switching functions⁹⁹.

A microcavity filter that has received considerable attention is one that enables resonant transfer of optical power between two waveguides. In its simplest form, it consists of a single whispering gallery microresonator sandwiched between two single-mode waveguides (Fig. 6). As a passive filter, this structure can perform a function called channel add/drop in which a single channel is 'dropped' with high extinction from a first waveguide and coupled with low loss to a second waveguide. In practice, this coupling should take place with high selectivity because other wavelength channels will be present on the input waveguide. Other versions of this device (that are not based on a microresonator) are also being investigated^{99–101}; however, the microresonator version is attractive because of its small size and potential for high-density integration on a wafer¹⁰². If the whispering gallery resonator is made active by addition of an electrically controllable refractive index, then the add/drop function is dynamic and provides control of the resonant wavelength for tuning¹⁰³, ultrafast modulation¹⁰⁴ or switching. Alternatively, introducing a controlled loss within the resonator enables switching-off of the coupling^{105–107}. Arrays of these devices on a common sub-

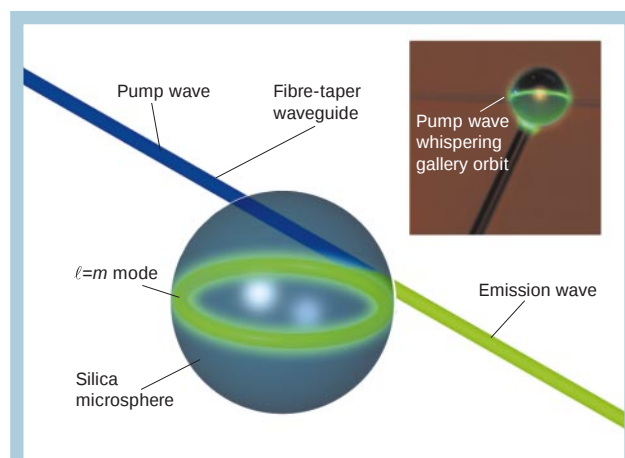


Figure 5 Illustration of a silica microsphere whispering gallery resonator. The green orbit is a $l=m$ mode entrained at the sphere's surface. Also shown is a fibre taper waveguide used for power coupling to and from the resonator. In the figure, a blue pump wave induces a circulating intensity within the sphere that is sufficient to induce laser oscillation (green emission wave). Inset: Photomicrograph showing a doped microsphere (the glass sphere contains a rare earth). The green emission in this case traces the pump wave whispering gallery orbit. The inset micrograph was provided by M. Cai.

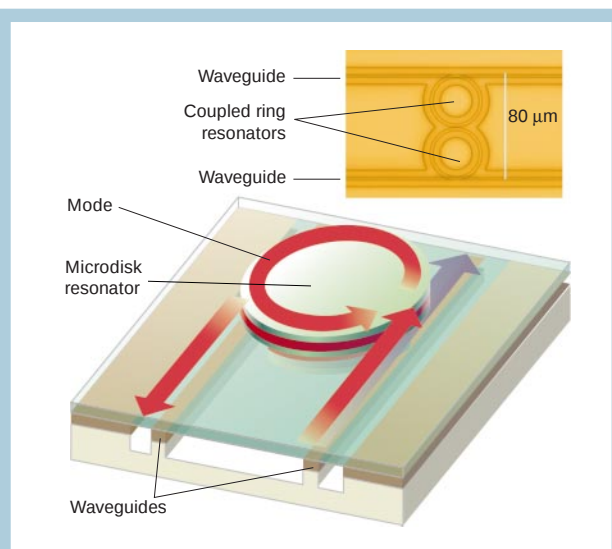


Figure 6 Illustration of a microcavity add/drop filter in which two buried waveguides (shown in brown) are vertically coupled to a disk whispering gallery resonator. This structure is fabricated by first creating the waveguides through a process of lithography and etching and then wafer bonding an initially mechanically separate, second wafer containing layers that ultimately become the microresonator. After wafer bonding, the substrate is removed from the second wafer, and lithography and etching are applied to create the microdisk. Critical gaps that control field-coupling strengths between waveguides and the microresonator are thereby determined by thin-film crystal growth methods as opposed to lithography and etching as is the case in side-coupled structures. As shown, a red input channel is resonant with a mode of the disk and is subsequently coupled through to the second waveguide. A blue channel that is non-resonant is also shown. The channel add function is not illustrated; however, it results from coupling power into the other port of the drop waveguide with subsequent resonant coupling to the input waveguide along the indicated direction of signal flow. This device can be made dynamic by inclusion of electrically controllable loss or refractive layers within the resonator. Low-loss coupling and high drop extinction require the microdisk to have an intrinsic Q factor, Q_0 , that is substantially larger than its Q factor under conditions of waveguide loading, Q_L . Furthermore, efficient coupling requires waveguide loading to be balanced, such that field coupling between the disk and each waveguide is nearly equal. Q_L is then determined by the requisite information bandwidth, B , with $B < \nu/Q_L$ required for proper information transfer. This schematic is based on one by P. D. Dapkus¹⁰³. The inset is a photomicrograph of a ring-resonator add/drop filter provided by Brent Little (Little Optics). The device shown is a 'second order' design containing a pair of coupled ring resonators. The waveguides (top and bottom) feature 'vertical coupling' to the resonators, providing excellent coupling control.

strate^{108,109} or interconnected by a fibre could one day be used to perform complex functions within the context of WDM.

Many versions of this basic design have been demonstrated, including monolithic III–V¹¹⁰, silica¹¹¹ and polymer-based devices¹⁰⁴. Arrays have also been fabricated to explore issues associated with fabrication tolerance and post-process trimming¹⁰². The major challenge for fabrication of these devices is ensuring control and reproducibility of waveguide-to-resonator coupling and resonator dimensions. One approach to obtaining better fabrication control of coupling is to use so-called vertical coupling structures (coupling regions that are defined primarily by layer thickness as opposed to lithography)^{104,106,111}.

Another challenge for fabrication concerns the spectral shape and filter roll-off characteristics of these devices. Spectral response functions with roll-offs that are steeper than are possible using a simple Lorentzian (single pole) filter response are necessary for concatenation of components in networks¹¹². Designs for such multi-pole filters based on coupled resonators¹¹³ have been proposed, and several devices have been demonstrated^{114,115}. Active frequency control¹⁰³ or

UV trimming of device refractive index¹⁰² could be required to offset fabrication-induced variances in critical resonator dimensions.

Future directions

We have reviewed four broad application areas of optical microcavities and highlighted several microcavity designs for each (see Table 1). Impressive results have been achieved in all areas. Substantial, additional gains are possible in quantum optical applications with continued improvement in microfabrication techniques and with implementation of new low-loss designs. Triggered, single photon sources will benefit from higher Purcell factors for improved fibre coupling, and miniaturization to the submicrometre scale of cavity QED devices (using either strong or weak coupling) is feasible. Also, the emergence of new ultrahigh- Q , wafer-based geometries should provide a platform for strong-coupling studies that combine both laboratory-on-chip functions and efficient coupling to optical fibres. Technological applications such as the dynamic add/drop device will provide better control and reproducibility of filter characteristics in designs that are increasingly complex.

One other area that deserves special note is that of biological and chemical sensing. Optical sensors that use evanescent field coupling have been developed^{116,117}; however, high- Q optical microcavities, as a sensor transducer, offer the potential to greatly enhance detection sensitivity³⁹. Recently, sensors based on both monolithic¹¹⁸ and microsphere¹¹⁹ whispering gallery transducers have been demonstrated. It seems likely that this will become an important application area for these devices. Likewise, the broad technological impact that resonant devices have had at acoustic, radio and microwave frequencies suggests that many other applications for these devices will emerge in the optical domain. □

doi:10.1038/nature01939

- Gerard, J. M. *et al.* Quantum boxes as active probes for photonic microstructures: The pillar microcavity case. *Appl. Phys. Lett.* **69**, 449–451 (1996).
- Solomon, G. S., Pelton, M. & Yamamoto, Y. Modification of spontaneous emission of a single quantum dot. *Phys. Status Solidi* **178**, 341–344 (2000).
- Jewell, J. L. *et al.* Lasing characteristics of GaAs microresonators. *Appl. Phys. Lett.* **54**, 1400–1402 (1989).
- Chang, R. K. (ed.) *Optical Processes in Microcavities* (World Scientific, Singapore, 1996).
- Rayleigh, L. in *Scientific Papers* 617–620 (Cambridge Univ., Cambridge, 1912).
- Armani, D. K., Kippenberg, T. J., Spillane, S. M. & Vahala, K. J. Ultra-high- Q toroid microcavity on a chip. *Nature* **421**, 925–928 (2003).
- Painter, O. *et al.* Two-dimensional photonic band-gap defect mode laser. *Science* **284**, 1819–1821 (1999).
- Taranenko, V. B. & Weiss, C. O. Spatial solitons in semiconductor microresonators. *IEEE J. Select. Top. Quantum Electron.* **8**, 488–496 (2002).
- Barland, S. *et al.* Cavity solitons as pixels in semiconductor microcavities. *Nature* **419**, 699–702 (2002).
- Stone, A. D. Wave-chaotic optical resonators and lasers. *Physica Scripta* **T90**, 248–262 (2001).
- Khitrova, G., Gibbs, H. M., Jahnke, F., Kira, M. & Koch, S. W. Nonlinear optics of normal-mode-coupling semiconductor microcavities. *Rev. Mod. Phys.* **71**, 1591–1639 (1999).
- Brinkman, W. F., Koch, T. L., Lang, D. V. & Wilt, D. P. The lasers behind the communications revolution. *Bell Labs Tech. J.* **5**, 150–167 (2000).
- Coldren, L. A. & Corzine, S. W. *Diode Lasers and Photonic Integrated Circuits* (Wiley, New York, 1995).
- Yariv, A. *Quantum Electronics* (Wiley, New York, 1989).
- Berman, P. R. (ed.) *Cavity Quantum Electrodynamics* (Academic Press, New York, 1993).
- Haroche, S. Entanglement, mesoscopic superpositions and decoherence studies with atoms and photons in a cavity. *Physica Scripta* **T76**, 159–164 (1998).
- Kimble, H. J. Strong interactions of single atoms and photons in cavity QED. *Physica Scripta* **T76**, 127–137 (1998).
- Thompson, R. J., Rempe, G. & Kimble, H. J. Observation of normal-mode splitting for an atom in an optical cavity. *Phys. Rev. Lett.* **68**, 1132–1135 (1992).
- Mabuchi, H., Turchette, Q. A., Chapman, M. S. & Kimble, H. J. Real-time detection of individual atoms falling through a high-finesse optical cavity. *Opt. Lett.* **21**, 1393–1395 (1996).
- Rempe, G. One-atom in an optical cavity - spatial-resolution beyond the standard diffraction limit. *Appl. Phys. B* **60**, 233–237 (1995).
- Hood, C. J., Lynn, T. W., Doherty, A. C., Parkins, A. S. & Kimble, H. J. The atom-cavity microscope: Single atoms bound in orbit by single photons. *Science* **287**, 1447–1453 (2000).
- Pinkse, P. W. H., Fischer, T., Maunz, P. & Rempe, G. Trapping an atom with single photons. *Nature* **404**, 365–368 (2000).
- Shimizu, Y. *et al.* Control of light pulse propagation with only a few cold atoms in a high-finesse microcavity. *Phys. Rev. Lett.* **89**, 233001 (2002).
- Mabuchi, H. & Doherty, A. C. Cavity quantum electrodynamics: Coherence in context. *Science* **298**, 1372–1377 (2002).
- Raimond, J. M., Brune, M. & Haroche, S. Colloquium: Manipulating quantum entanglement with atoms and photons in a cavity. *Rev. Mod. Phys.* **73**, 565–582 (2001).
- Rempe, G., Thompson, R. J., Kimble, H. J. & Lalezari, R. Measurement of ultralow losses in an optical interferometer. *Opt. Lett.* **17**, 363–365 (1992).

27. Lefevre-Seguin, V. & Haroche, S. Towards cavity-QED experiments with silica microspheres. *Mat. Sci. Eng. B* **48**, 53–58 (1997).
28. Gorodetsky, M. L., Savchenkov, A. A. & Ilchenko, V. S. Ultimate Q of optical microsphere resonators. *Opt. Lett.* **21**, 453–455 (1996).
29. Vernooy, D. W., Ilchenko, V. S., Mabuchi, H., Streed, E. W. & Kimble, H. J. High-Q measurements of fused-silica microspheres in the near infrared. *Opt. Lett.* **23**, 247–249 (1998).
30. Vuckovic, J., Loncar, M., Mabuchi, H. & Scherer, A. Design of photonic crystal microcavities for cavity QED. *Phys. Rev. E* **65**, 016608 (2002).
31. Buck, J. R. & Kimble, H. J. Optimal sizes of dielectric microspheres for cavity QED with strong coupling. *Phys. Rev. A* **68**, 033806 (2003).
32. Hood, C. J., Kimble, H. J. & Ye, J. Characterization of high-finesse mirrors: Loss, phase shifts, and mode structure in an optical cavity. *Phys. Rev. A* **64**, 033804 (2001).
33. Knight, J. C. *et al.* Mapping whispering-gallery modes in microspheres with a near-field probe. *Opt. Lett.* **20**, 1515–1517 (1995).
34. Vernooy, D. W., Furusawa, A., Georgiades, N. P., Ilchenko, V. S. & Kimble, H. J. Cavity QED with high-Q whispering gallery modes. *Phys. Rev. A* **57**, R2293–R2296 (1998).
35. Braginsky, V. B., Gorodetsky, M. L. & Ilchenko, V. S. Quality-factor and nonlinear properties of optical whispering-gallery modes. *Phys. Lett. A* **137**, 393–397 (1989).
36. Gorodetsky, M. L. & Ilchenko, V. S. Optical microsphere resonators: optimal coupling to high-Q whispering-gallery modes. *J. Opt. Soc. Am. B* **16**, 147–154 (1999).
37. Ilchenko, V. S., Yao, X. S. & Maleki, L. Pigtailling the high-Q microsphere cavity: a simple fiber coupler for optical whispering-gallery modes. *Opt. Lett.* **24**, 723–725 (1999).
38. Treussart, F. *et al.* Microspheres based on silica microspheres. *Ann. Telecommun.* **52**, 557–568 (1997).
39. Serpenguzel, A., Arnold, S. & Griffel, G. Excitation of resonances of microspheres on an optical fiber. *Opt. Lett.* **20**, 654–656 (1995).
40. Knight, J. C., Cheung, G., Jacques, F. & Birks, T. A. Phase-matched excitation of whispering-gallery-mode resonances by a fiber taper. *Opt. Lett.* **22**, 1129–1131 (1997).
41. Cirac, J. I., van Enk, S. J., Zoller, P., Kimble, H. J. & Mabuchi, H. Quantum communication in a quantum network. *Physica Scripta* **T76**, 223–232 (1998).
42. Cai, M., Painter, O. & Vahala, K. J. Observation of critical coupling in a fiber taper to a silica-microsphere whispering-gallery mode system. *Phys. Rev. Lett.* **85**, 74–77 (2000).
43. Spillane, S. M., Kippenberg, T. J. & Vahala, K. J. Ultralow-threshold Raman laser using a spherical dielectric microcavity. *Nature* **415**, 621–623 (2002).
44. Purcell, E. M. Spontaneous emission probabilities at radio frequencies. *Phys. Rev.* **69**, 681–681 (1946).
45. Yamamoto, Y., Tassone, F. & Cao, H. *Semiconductor Cavity Quantum Electrodynamics* (Springer, New York, 2000).
46. Haroche, S. & Kleppner, D. Cavity quantum electrodynamics. *Phys. Today* **42**, 24–30 (1989).
47. Marzin, J. Y., Gerard, J. M., Israel, A., Barrier, D. & Bastard, G. Photoluminescence of single InAs quantum dots obtained by self-organized growth on GaAs. *Phys. Rev. Lett.* **73**, 716–719 (1994).
48. Gerard, J. M. *et al.* Enhanced spontaneous emission by quantum boxes in a monolithic optical microcavity. *Phys. Rev. Lett.* **81**, 1110–1113 (1998).
49. Gerard, J. M. & Gayral, B. InAs quantum dots: artificial atoms for solid-state cavity-quantum electrodynamics. *Physica E* **9**, 131–139 (2001).
50. Gerard, J. M. *et al.* InAs quantum boxes in GaAs/AlAs pillar microcavities: From spectroscopic investigations to spontaneous emission control. *Physica E* **2**, 804–808 (1998).
51. Bayer, M. *et al.* Inhibition and enhancement of the spontaneous emission of quantum dots in structured microresonators. *Phys. Rev. Lett.* **86**, 3168–3171 (2001).
52. Gayral, B. *et al.* High-Q wet-etched GaAs microdisks containing InAs quantum boxes. *Appl. Phys. Lett.* **75**, 1908–1910 (1999).
53. Michler, P. *et al.* Quantum dot lasers using high-Q microdisk cavities. *Phys. Status Solidi B* **224**, 797–801 (2001).
54. Kiraz, A. *et al.* Cavity-quantum electrodynamics using a single InAs quantum dot in a microdisk structure. *Applied Phys. Lett.* **78**, 3932–3934 (2001).
55. Gerard, J. M. & Gayral, B. Strong Purcell effect for InAs quantum boxes in three-dimensional solid-state microcavities. *J. Lightwave Technol.* **17**, 2089–2095 (1999).
56. Pelton, M., Vuckovic, J., Solomon, G. S., Scherer, A. & Yamamoto, Y. Three-dimensionally confined modes in microphotonic microcavities: Quality factors and Purcell factors. *IEEE J. Quantum Elect.* **38**, 170–177 (2002).
57. Vuckovic, J., Pelton, M., Scherer, A. & Yamamoto, Y. Optimization of three-dimensional microphotonic microcavities for cavity quantum electrodynamics. *Phys. Rev. A* **66**, 023808 (2002).
58. Yablonovitch, E. Inhibited spontaneous emission in solid-state physics and electronics. *Phys. Rev. Lett.* **58**, 2059–2062 (1987).
59. Reese, C. *et al.* High-Q photonic crystal microcavities fabricated in a thin GaAs membrane. *J. Vac. Sci. Technol. B* **19**, 2749–2752 (2001).
60. Happ, T. D. *et al.* Enhanced light emission of InxGa1-xAs quantum dots in a two-dimensional photonic-crystal defect microcavity. *Phys. Rev. B* **66**, 041303 (2002).
61. Srinivasan, K. & Painter, O. Momentum space design of high-Q photonic crystal optical cavities. *Opt. Exp.* **10**, 670–684 (2002).
62. Srinivasan, K., Barclay, P., Painter, O., Chen, J., Cho, C. & Gmachl, C. Experimental demonstration of a high quality factor photonic crystal microcavity. *Appl. Phys. Lett.* (in the press).
63. Boroditsky, M. *et al.* Spontaneous emission extraction and Purcell enhancement from thin-film 2-D photonic crystals. *J. Lightwave Technol.* **17**, 2096–2112 (1999).
64. D. Bouwmeester, A. E. & A. Zeilinger. *The Physics of Quantum Information* (Springer, Berlin, 2000).
65. Kimble, H. J., Dagenais, M. & Mandel, L. Photon anti-bunching in resonance fluorescence. *Phys. Rev. Lett.* **39**, 691–695 (1977).
66. Basche, T., Moerner, W. E., Orrit, M. & Talon, H. Photon antibunching in the fluorescence of a single dye molecule trapped in a solid. *Phys. Rev. Lett.* **69**, 1516–1519 (1992).
67. Michler, P. *et al.* Quantum correlation among photons from a single quantum dot at room temperature. *Nature* **406**, 968–970 (2000).
68. Michler, P. *et al.* A quantum dot single-photon turnstile device. *Science* **290**, 2282–2285 (2000).
69. Santori, C., Pelton, M., Solomon, G., Dale, Y. & Yamamoto, E. Triggered single photons from a quantum dot. *Phys. Rev. Lett.* **86**, 1502–1505 (2001).
70. Zwiller, V. *et al.* Single quantum dots emit single photons at a time: Antibunching experiments. *Appl. Phys. Lett.* **78**, 2476–2478 (2001).
71. Thompson, R. M. *et al.* Single-photon emission from exciton complexes in individual quantum dots. *Phys. Rev. B* **64**, 201302 (2001).
72. Lounis, B. & Moerner, W. E. Single photons on demand from a single molecule at room temperature. *Nature* **407**, 491–493 (2000).
73. Brunel, C., Lounis, B., Tamarat, P. & Orrit, M. Triggered source of single photons based on controlled single molecule fluorescence. *Phys. Rev. Lett.* **83**, 2722–2725 (1999).
74. Moreau, E. *et al.* Single-mode solid-state single photon source based on isolated quantum dots in pillar microcavities. *Appl. Phys. Lett.* **79**, 2865–2867 (2001).
75. Santori, C., Fattal, D., Vuckovic, J., Solomon, G. S. & Yamamoto, Y. Indistinguishable photons from a single-photon device. *Nature* **419**, 594–597 (2002).
76. Pelton, M. *et al.* Efficient source of single photons: A single quantum dot in a microphotonic microcavity. *Phys. Rev. Lett.* **89**, 233602 (2002).
77. Yuan, Z. L. *et al.* Electrically driven single-photon source. *Science* **295**, 102–105 (2002).
78. Qian, S. X., Snow, J. B., Zheng, H. M. & Chang, R. K. Lasing droplets - highlighting the liquid-air interface by laser-emission. *Science* **231**, 486–488 (1986).
79. Lin, H. B., Eversole, J. D. & Campillo, A. J. Spectral properties of lasing microdroplets. *J. Opt. Soc. Am. B* **9**, 43–50 (1992).
80. Cai, M., Painter, O., Vahala, K. J. & Sercel, P. C. Fiber-coupled microsphere laser. *Opt. Lett.* **25**, 1430–1432 (2000).
81. Kuwatagano, M., Takeda, K., Yasuda, H. & Ema, K. Laser-emission from dye-doped polystyrene microsphere. *Japanese J. Appl. Phys.* **31**, L99–L101 (1992).
82. McCall, S. L., Levi, A. F. J., Slusher, R. E., Pearson, S. J. & Logan, R. A. Whispering-gallery mode microdisk lasers. *Appl. Phys. Lett.* **60**, 289–291 (1992).
83. Rex, N. B., Chang, R. K. & Guido, L. J. Threshold lowering in GaN micropillar lasers by means of spatially selective optical pumping. *IEEE Phot. Technol. Lett.* **13**, 1–3 (2001).
84. Deppe, D. G., Huffaker, D. L., Oh, T. H., Deng, H. Y. & Deng, Q. Low-threshold vertical-cavity surface-emitting lasers based on oxide-confinement and high contrast distributed Bragg reflectors. *IEEE J. Select. Top. Quantum Elect.* **3**, 893–904 (1997).
85. Sandoghdar, V. *et al.* Very low threshold whispering-gallery-mode microsphere laser. *Phys. Rev. A* **54**, R1777–R1780 (1996).
86. Macdougall, M. H., Dapkus, P. D., Pudikov, V., Zhao, H. M. & Yang, G. M. Ultralow threshold current vertical-cavity surface-emitting lasers with AlAs oxide-GaAs Distributed Bragg reflectors. *IEEE Phot. Technol. Lett.* **7**, 229–231 (1995).
87. Huffaker, D. L., Graham, L. A., Deng, H. & Deppe, D. G. Sub-40 microAmp continuous-wave lasing in an oxidized vertical-cavity surface-emitting laser with dielectric mirrors. *IEEE Phot. Technol. Lett.* **8**, 974–976 (1996).
88. Yamamoto, Y. & Slusher, R. E. Optical processes in microcavities. *Phys. Today* **46**, 66–73 (1993).
89. Rice, P. B. & Carmichael, H. J. Photon statistics of a cavity-QED laser - a comment on the laser-phase-transition analogy. *Phys. Rev. A* **50**, 4318–4329 (1994).
90. An, K. & Feld, M. A. Semiclassical four-level single atom laser. *Phys. Rev. A* **56**, 1662–1665 (1997).
91. Graham, L. A., Huffaker, D. L., Csutak, S. M., Deng, Q. & Deppe, D. G. Spontaneous lifetime control of quantum dot emitters in apertured microcavities. *J. Appl. Phys.* **85**, 3383–3385 (1999).
92. Qian, S. X., Snow, J. B. & Chang, R. K. Coherent Raman mixing and coherent anti-Stokes Raman-scattering from individual micrometer-size droplets. *Opt. Lett.* **10**, 499–501 (1985).
93. Lin, H. B., Eversole, J. D. & Campillo, A. J. Continuous-wave stimulated Raman-scattering in microdroplets. *Opt. Lett.* **17**, 828–830 (1992).
94. Qian, S. X. & Chang, R. K. Multiorder Stokes emission from micrometer-size droplets. *Phys. Rev. Lett.* **56**, 926–929 (1986).
95. Zhang, J. Z., Chen, G. & Chang, R. K. Pumping of stimulated Raman-scattering by stimulated Brillouin-scattering within a single liquid droplet - input laser linewidth effects. *J. Opt. Soc. Am. B* **7**, 108–115 (1990).
96. Treussart, F. *et al.* Evidence for intrinsic Kerr bistability of high-Q microsphere resonators in superfluid helium. *Eur. Phys. J. D* **1**, 235–238 (1998).
97. Kaminow, I. P. & San, T. L. (eds) *Optical Fiber Telecommunications IV* (Academic Press, San Diego, 2002).
98. Alexander, S. B. *et al.* A precompetitive consortium on wide-band all-optical networks. *J. Lightwave Technol.* **11**, 714–735 (1993).
99. Hibino, Y., Maruno, T. & Okamoto, K. Recent progress on large-scale PLC technologies with advanced functions. *NTT Review* **13**, 4–9 (2001).
100. Kewitsch, A. S., Rakuljic, G. A., Willems, P. A. & Yariv, A. All-fiber zero-insertion-loss add-drop filter for wavelength-division multiplexing. *Opt. Lett.* **23**, 106–108 (1998).
101. Doerr, C. R. *et al.* 40-wavelength add-drop filter. *IEEE Phot. Technol. Lett.* **11**, 1437–1439 (1999).
102. Suzuki, S., Hatakeyama, Y., Kokubun, Y. & Chu, S. T. Precise control of wavelength channel spacing of microring resonator add-drop filter array. *J. Lightwave Technol.* **20**, 745–750 (2002).
103. Djordjevic, K., Choi, S. J. & Dapkus, P. D. Microdisk tunable resonant filters and switches. *IEEE Phot. Technol. Lett.* **14**, 828–830 (2002).
104. Rabiei, P., Steier, W. H., Cheng Zhang & Dalton, L. R. Polymer micro-ring filters and modulators. *J. Lightwave Technol.* **20**, 1968–1975 (2002).
105. Little, B. E. *et al.* Wavelength switching and routing using absorption and resonance. *IEEE Phot. Technol. Lett.* **10**, 816–818 (1998).
106. Djordjevic, K., Choi, S. J. & Dapkus, P. D. Vertically coupled InP microdisk switching devices with electroabsorptive active regions. *IEEE Phot. Technol. Lett.* **14**, 1115–1117 (2002).
107. Yariv, A. S. Critical coupling and its control in optical waveguide-ring resonator systems. *IEEE Phot. Technol. Lett.* **14**, 483–485 (2002).
108. Soref, R. A. & Little, B. E. Proposed N-wavelength M-fiber WDM crossconnect switch using active microring resonators. *IEEE Phot. Technol. Lett.* **10**, 1121–1123 (1998).
109. Chu, S. T. *et al.* An eight-channel add-drop filter using vertically coupled microring resonators over a cross grid. *IEEE Phot. Technol. Lett.* **11**, 691–693 (1999).
110. Grover, R. *et al.* Vertically coupled GaInAsP-InP microring resonators. *Opt. Lett.* **26**, 506–508 (2001).
111. Little, B. E. *et al.* Vertically coupled glass microring resonator channel dropping filters. *IEEE Phot. Technol. Lett.* **11**, 215–217 (1999).
112. Offrein, B. J. *et al.* Resonant coupler-based tunable add-after-drop filter in silicon-oxynitride technology for WDM networks. *IEEE J. Select. Top. Quantum Elect.* **5**, 1400–1406 (1999).
113. Little, B. E., Chu, S. T., Haus, H. A., Foresi, J. & Laine, J. P. Microring resonator channel dropping filters. *J. Lightwave Technol.* **15**, 998–1005 (1997).

114. Grover, R. *et al.* Parallel-cascaded semiconductor microring resonators for high-order and wide-FSR filters. *J. Lightwave Technol.* **20**, 872–877 (2002).
115. Yanagase, Y., Suzuki, S., Kokubun, Y. & Chu, S. T. Box-like filter response and expansion of FSR by a vertically triple coupled microring resonator filter. *J. Lightwave Technol.* **20**, 1525–1529 (2002).
116. Cooper, M. A. Optical biosensors in drug discovery. *Nat. Rev. Drug Discov.* **1**, 515–527 (2002).
117. Marazuela, M. D. & Moreno-Bondi, M. C. Fiber-optic biosensors - an overview. *Anal. Bioanal. Chem.* **372**, 664–682 (2002).
118. Krioukov, E., Klunder, D. J. W., Driessen, A., Greve, J. & Otto, C. Sensor based on an integrated optical microcavity. *Opt. Lett.* **27**, 512–514 (2002).
119. Vollmer, F. *et al.* Protein detection by optical shift of a resonant microcavity. *Appl. Phys. Lett.* **80**, 4057–4059 (2002).
120. Walther, H. Quantum optics of a single atom. *Physica Scripta* **1**, 138–146 (1998)

Acknowledgements This work was supported by DARPA, the Caltech Lee Centre and the National Science Foundation.

Photonic crystal fibres

Jonathan C. Knight

Department of Physics, University of Bath, Claverton Down, Bath, BA2 7AY, UK (e-mail: j.c.knight@bath.ac.uk)

Photonic crystal fibres have wavelength-scale morphological microstructure running down their length. This structure enables light to be controlled within the fibre in ways not previously possible or even imaginable. Our understanding of what an optical fibre is and what it does is changing because of the development of this new technology, and a broad range of applications based on these principles is being developed.

Could optical fibres as we now know them become obsolete? It seems an unlikely proposition, and yet a new generation of optical fibres currently being developed could outperform conventional fibres for many applications. Modern optical fibres were one of the major technological successes of the 20th century, and they have become the pre-eminent method of communicating information. This technology has developed at an incredible pace, from the first low-loss (<20 dB/km) single-mode waveguides in 1970 to being key components of our sophisticated global telecommunication network. Modern optical fibres, which transmit information in the form of short optical pulses over long distances at exceptionally high speeds, have become an integral part of life in the information age. There are non-telecom applications for fibres, too, in beam delivery for medicine, machining and diagnostics, sensing and a host of other fields. And yet, amazingly, the basic physics of optical fibres has remained unchanged since the 19th century. This, in part, may be a direct result of the fast initial implementation of fibre optic technology. The system was established early on, and all future developments were incremental improvements to individual components. State-of-the-art optical fibres represent a careful trade-off between optical losses, optical nonlinearity, group-velocity dispersion and polarization effects. Some of these effects (e.g. optical losses) are inherent in the raw material used to make the fibres, which is usually synthetically produced silica, SiO₂. Others (nonlinearity and dispersion) are strongly affected by the material's properties but can be influenced by fibre design. Other properties such as polarization-mode dispersion result from imperfections in the fabrication processes. After 30 years of intensive research, incremental steps have refined the capabilities of the system and the fabrication technology nearly as far as they can go.

Researchers and engineers in several laboratories around the world are working on ways to revolutionize fibre-optic design and performance. Since the 1980s, optical physicists have recognized that the ability to structure materials on the scale of the optical wavelength, a fraction of a micrometre or less, will allow the development of new optical materials known as photonic crystals. Photonic crystals rely on regular morphological microstructure incorporated into the material to radically alter its optical properties. Many research groups are investigating these materials in two and three dimensional and planar geometries, and a few researchers are using them to form new types of optical fibre. Such fibres are known as photonic crystal fibres (PCFs), as they rely on the unusual properties of photonic crystals to deliver previously unimaginable performance from an optical fibre waveguide. To be competitive with existing telecommunication

technology, new fibres would have to perform at least as well as conventional fibres overall and deliver significantly improved performance in some respects. For non-telecom applications, simply being superior in one or a few characteristics could make the alternative fibre technologies valuable. PCFs are already demonstrating such superiority in several respects, which is leading to new applications. They have the potential to outperform conventional fibres in other areas as well, and they are changing our understanding of what an optical fibre is.

This article gives a qualitative description of the basic physics of the new fibres, and a snapshot of the current state of the art. Many of the unanswered questions about the technology can only be addressed experimentally, and the concerted efforts of the past few years have resulted in impressive progress. The potential for photonic crystals remains huge but largely unproven, but we expect to establish some of the limitations of the new fibres within the next few years.

Optics in a fibre

Optical fibres are an excellent approximation of two-dimensional (2D) structures because they are effectively infinite in the third dimension. They are nominally invariant along their length, with all interfaces parallel to the fibre axis. Propagation of light in such a structure is best characterized by remembering that when light with free-space propagation constant k encounters an interface between two materials with refractive indices n_1 and n_2 , the component of the wavevector parallel to the interface remains unchanged during the interaction. This rule is most widely known (Fig. 1a) through the law of reflection (that the angle of incidence is equal to the angle of reflection, so that the quantity $nk \sin \theta$ is preserved in reflection) and the law of refraction (that $n_1 \sin \theta_1 = n_2 \sin \theta_2$, so that $nk \sin \theta$ is preserved in transmission as well). In fibre optics, this is so important that we parameterize the problem in terms of β (as shown in Fig. 1b), the component of the propagation constant along the fibre axis, as this is a constant for a given mode of propagation. Because the only interfaces in a fibre that are not parallel to β are the input and output faces, light introduced into the fibre with a given value of β should maintain that value until it leaves the fibre, whether it is in a trapped core mode or a mode of the cladding.

To form a guided mode in the core, one needs to introduce light into the core with a value of β that cannot propagate in the cladding. The largest value of β that can exist in an infinite homogeneous medium with refractive index n is $\beta = nk$, with all smaller values of β allowed. We derive from β a modal index $n_m = \beta/k$. The range of modal indices supported by specific materials—be they conventional homogeneous material or artificially fabricated photonic crystals—can then be used to identify ways in which they can be used to form guided modes (Fig. 1c). As with any material, a 2D photonic crystal has a maximum β value that can propagate. At a particular frequency,

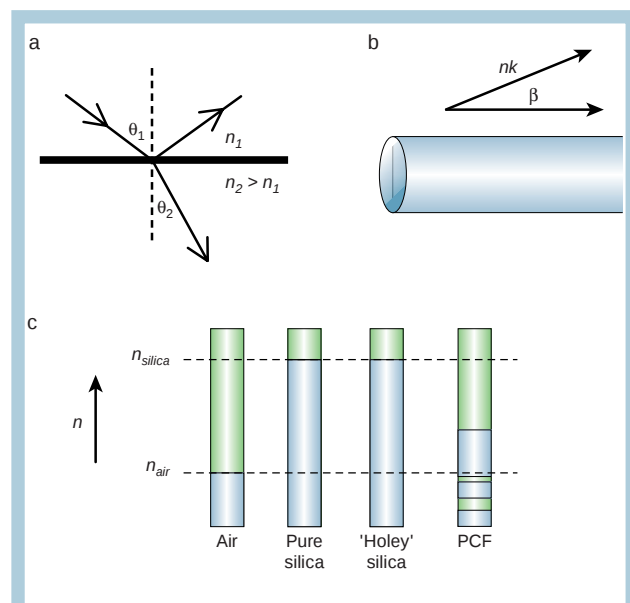


Figure 1 Propagation of light in an optical fibre. **a**, Reflection and refraction at an interface. In reflection, the angle of incidence is equal to the angle of reflection. In refraction, the angles obey $n_1 \sin \theta_1 = n_2 \sin \theta_2$. In both cases, the quantity $\beta = nk \sin \theta$ is preserved. **b**, Definition of β for propagation in a fibre. The total propagation constant for light of wavelength λ in a medium of index n is nk , where $k = 2\pi/\lambda$ is the free-space wavevector. β is the component of the total wavevector that is parallel to the fibre axis. **c**, The allowed values (blue) of modal index for (i) air, (ii) pure silica, (iii) 'holey' silica with a reduced effective refractive index and (iv) a PCF structure with bandgaps for $\beta < k$. The dotted lines n_{air} and n_{silica} indicate the refractive indices of air and silica respectively.

this corresponds to the 'fundamental' mode of an infinite slab of the material, and this value of β defines the 'effective refractive index' of the material¹. Consequently, one can use a 2D photonic crystal as a fibre cladding at a particular frequency if a core material that has a higher refractive index than this effective index is chosen (Fig. 1c). An example of such a structure is a solid silica core surrounded by silica–air photonic crystal cladding. Such fibres will guide light (Fig. 2b) through a form of total internal reflection (TIR); nonetheless, in many respects they represent a radical departure from conventional fibre optics. However, far more radical fibre designs result from the fact that the range of modal indices supported by photonic crystals have gaps in them—ranges of modal index where there are no propagating modes—despite the existence of such modes for larger and smaller indices (Fig. 1c). These are the photonic bandgaps of the crystal, which are similar to the 2D bandgaps that are being investigated for planar lightwave circuits, but here they have propagation with a non-zero value of β (ref. 2). Accurate plots similar to those in Fig. 1c can be created by performing detailed photonic band structure calculations to determine which modal frequencies are supported at a given value of β . As with 3D and in-plane 2D band gaps, out-of-plane bandgaps can be thought of as arising

when Bragg scattering is extended to a strongly scattering environment or due to coupling within an array of resonators. The bandgaps in this case appear as functions of β rather than of optical frequency. Crucially, gaps can appear for values of modal index both greater and smaller than unity, enabling the formation of hollow-core fibres with bandgap material as a cladding (Figs 2c and 3a, d). Such fibres are related to Bragg fibres³ in that they do not rely on TIR to guide light.

Fabrication and materials

A key factor, perhaps the key factor, in designing and developing new fibres is their fabrication: what material systems can be used, and what technologies are required for their development? Whereas producing conventional single-mode optical fibre requires core and cladding materials with similar refractive indices (typically differing by around a percent), designing photonic crystal fibres requires a far higher refractive index contrast, differing by perhaps 50–100%. The small index differences in conventional fibres are usually achieved by vapour deposition techniques, but the larger index contrasts needed for PCF have been most widely attained by developing⁴ the tube-stacking technique investigated in the 1970s by Bell Laboratories⁵ and others or by directly forming the glass by using, for example, extrusion through a die^{6–8}. Recently, hollow-core concentric ring structures were formed from two solid materials⁹. Fibres resulting from these procedures will revolutionize fibre optics, by providing new tools with which to influence fibre properties while freeing the fibre from the constraints of naturally occurring materials.

In many of the studies reported to date, fibres have been formed using a solid glassy material pierced with air holes. This has several advantages: air is mechanically and thermally compatible with most materials, it is transparent over a broad spectral range, and it has a very low refractive index at optical frequencies. Designs of this type are restricted to those in which the second material—the high-index material—is connected. This restriction is significant for bandgap fibres, as weakly coupled high-index regions can form good bandgaps. Fibres formed using silica and air have been thoroughly investigated^{4,6,10}, partly because most conventional optical fibres are produced from fused silica. It is also an excellent material to work with because viscosity does not change much with temperature and it is relatively cheap.

To fabricate a PCF, one must first create a preform, which contains the structure of interest but on a macroscopic scale. This is then drawn down on a fibre-drawing tower, greatly extending its length while reducing its cross-section. For silica-based PCFs, the preform is usually made by stacking capillaries and rods by hand on a macroscopic scale: a typical preform would be a metre long, 20 mm in diameter, and contains several hundreds of capillaries. By jacketing this stack, and perhaps using pressure to maintain the air hole size in the fibre, a wide range of different structures have been made, each with different optical properties. Silica–air preforms have also been extruded⁶, enabling the formation of structures not readily attainable by stacking capillaries. The extrusion process has recently been applied to other glasses^{7,8} (Fig. 4c), which are not as readily available in tube form as silica. Filling the holes of a silica–air structure opens up a broad range of possibilities: as well as forming a range of different telecommunication devices¹¹, a filled fibre can exhibit bandgap

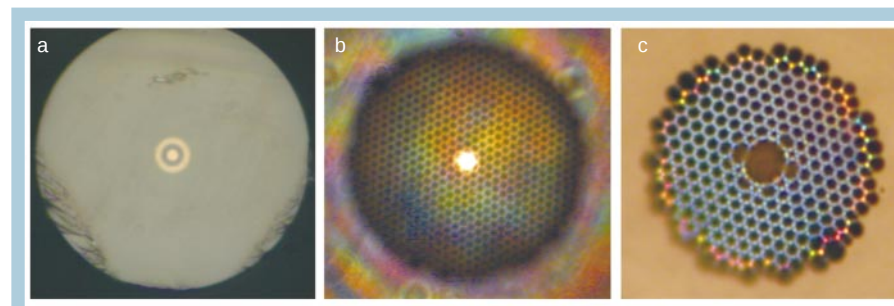


Figure 2 Conventional fibre and PCFs. Optical micrographs of (a) a standard optical fibre, formed using two bulk materials (core diameter 9 μm), (b) an index-guiding photonic crystal fibre (core diameter 5 μm) and (c) a hollow-core photonic bandgap fibre (core diameter 7 μm). The fibre in **a** is single-mode in the infrared but guides several modes at visible wavelengths. The fibre in **b** guides light in a single mode over the whole optical spectrum and that in **c** guides light only over a limited range of wavelengths in the infrared that correspond to the bandgap of the cladding material.

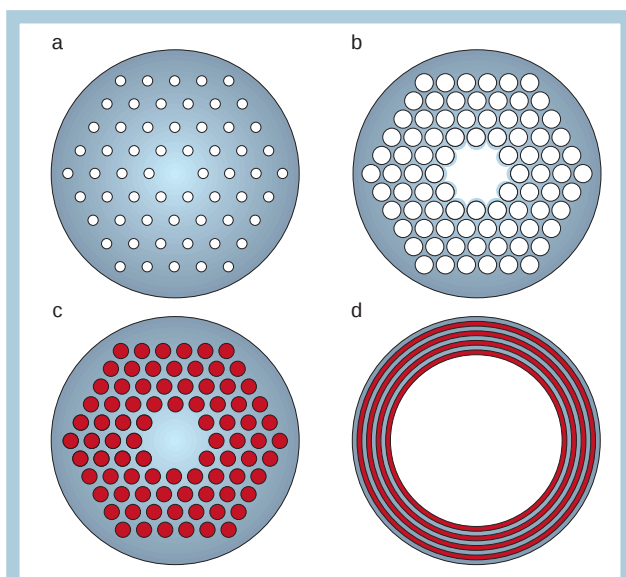


Figure 3 Different designs of PCF waveguide. **a** shows a single-mode fibre with a pure silica core surrounded by a reduced-index photonic crystal cladding material. The guidance is by modified total internal reflection. **b** is an air-guiding fibre in which light is confined to a hollow core by the bandgap of the 2D air-glass photonic crystal cladding. In **c**, the light is confined to a low-index region by a photonic bandgap, but the core is made of pure silica while the holes in the cladding are filled with a high-index liquid. **d** shows a hollow cylindrical multilayer fibre with an all-solid cladding. In these pictures, white represents air, blue represents a low-index solid such as silica and red represents a high-index material.

guidance in a low-index (pure silica) core when the holes are filled with a high-index liquid¹² (Fig. 3c). Stacking, extrusion and drilling are all means of forming PCFs from polymer and air¹³. A unique cigar-rolling technique has also been reported for a polymer-glass combination⁹. In this technique, a multilayer mirror is effectively rolled up to form a preform with a hollow core (Fig. 3d). This structure differs from the others reported above in several respects: it uses two solid materials but in a configuration that results in an almost exclusively radial variation in refractive index. The radial-only index variation has intrinsic advantages for forming hollow-core fibres: bandgap waveguiding in a hollow core is much easier to create, in principle, because only a single periodicity is involved. This makes it possible to form air-guided modes using two materials that would not provide an air-guiding bandgap when disposed in a 2D array. By contrast, the use of two solid materials, as required for a self-supporting radial-only variation, limits the materials that can be used to those with compatible thermal and thermo-mechanical properties. Excellent progress has been demonstrated to date⁹, and a structure with a bandgap in the 10 μm wavelength band has been produced. In that wavelength range, the fibre performance should be compared with chalcogenide fibres and inner-surface coated capillaries, which have so far performed relatively poorly. To reproduce the success of the hollow-core fibres at wavelengths shorter than 2 μm —where most laser sources and all of current optical telecommunications exist—will require further advances in fabrication procedures and perhaps new material systems.

The most important factor for any optical fibre technology is loss. Losses in conventional optical fibres have been reduced over the past 30 years, and, although they may fall further, they are unlikely to fall far. The minimum-loss wavelength in fused silica is around 1,550 nm, and the current loss levels are less than 0.2 dB/km at that wavelength. This number is important because it sets the amplifier spacing in long-haul communications systems, and thus is a major cost of a long-haul transmission system. In conventional silica fibres, losses increase

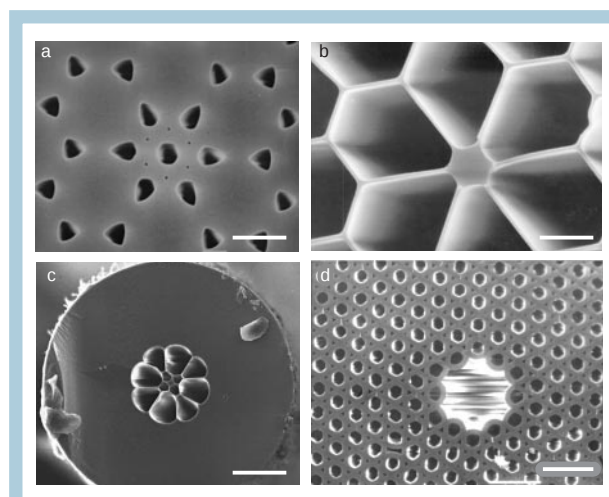


Figure 4 PCFs from the University of Bath. Electron micrographs of **(a)** bandgap-guiding fibre in which light is trapped in a ring of glass around a central hole (bar, 10 μm), **(b)** a strongly TIR guiding PCF with zero dispersion around 800 nm wavelength (bar, 1 μm), **(c)** extruded fibre formed from commercial SF6 glass (bar, 10 μm) and **(d)** hollow-core photonic bandgap fibre (bar, 10 μm).

towards shorter wavelengths because of the strong dependence of Rayleigh scattering (scattering from density fluctuations frozen into the material) on wavelength. With longer wavelengths, losses increase because of the intrinsic infrared absorption of the material. Losses of a fraction of a dB/km have been reported from solid-core PCF¹⁴, limited in this case by surface scattering from around the air holes. However, solid-core PCF is not expected to have significantly lower losses than conventional fibres. In hollow-core fibre, the lowest reported loss is 13 dB/km¹⁵ (Fig. 5), and in this case the fundamental loss limits are not yet established.

Total internal reflection waveguides

Light can be guided in PCFs by embedding a region of solid glass within an array of air holes. This approach has several important applications, and an excellent introduction is available in a recent review¹⁶. Index-guiding PCFs offer a wealth of new opportunities, both to those interested in fundamental fibre optics and to those looking to extend its applications. These opportunities stem from just a few special properties of the photonic crystal cladding, which are caused by the large refractive index contrast and the 2D nature of the microstructure. These affect the dispersion, the smallest attainable core size, the number of guided modes, the numerical aperture and the birefringence. For example, group velocity dispersion can be radically affected by pure waveguide dispersion¹⁷ in fibres with small cores and large air holes in the cladding (Fig. 4b). It can also be more subtly engineered¹⁸, using the dispersion of the photonic crystal material itself, so as to cancel the combined effects of the bulk silica dispersion and the waveguide dispersion over a broad wavelength range¹⁹. Such dispersion engineering enables the generation of a single-mode broadband optical supercontinuum²⁰, which is a spectacular light source that is already being used for frequency metrology²¹ and optical coherence tomography²². Other properties of the fibres can be engineered: one can create structures with enormous birefringence²³ or that support only a single mode that is independent of the wavelength^{1,4} or scale²⁴ of the structure. Such structures are leading to novel sensors^{25,26}, high-power fibre lasers²⁷ and to developments in other research fields.

Photonic bandgaps and hollow-core fibres

Fibres with hollow cores cannot be made using conventional optic fibres. This is because guidance by TIR requires a lower-index cladding material surrounding the core, and there are no suitable

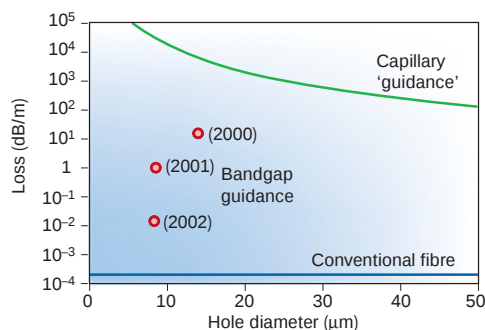


Figure 5 Predicted loss from the fundamental 'mode' in a silica capillary and reported losses from photonic bandgap fibres with hollow cores in the years 2000, 2001 and 2002. The bottom line indicates the current performance of conventional low-loss telecommunications fibre. On this scale, the lowest loss reported for index-guiding PCF is virtually indistinguishable from that for conventional fibres.

low-loss materials with a lower refractive index than air at optical frequencies. One can use a capillary: not a waveguide, but something that will help to confine the light over a certain propagation distance. Indeed, capillaries were investigated as a candidate for optical telecommunications in the 1960s²⁸, before the development of modern silica fibre technology. Another form of hollow core fibre is also in use: thin capillaries coated on their inner surface to increase their reflectance^{29,30}. This technology has been most widely (and usefully) applied at infrared wavelengths because the fabrication is much easier for the larger bore size required and because the coatings are easier to apply. With all these systems, however, the fibre lengths are relatively short (because the coating is applied after fibre drawing) and the modal quality is poor.

For PCFs, the fibre-drawing process reduces the transverse dimensions of the structured preform down to optical dimensions, and no further processing is required. This technique produces long lengths of fibre and allows the number of guided modes to be controlled through adjustment of the final core size. The first 'bandgap-guiding' PCF³¹ took the form of a close-packed stack of fine silica strands (Fig. 4a), but this fibre could only guide light in glass, the higher-index material. More recently, a pure silica core fibre surrounded by a lattice of holes filled with a higher-index liquid was reported¹² (Fig. 3c). This interesting structure has isolated islands of high-index material embedded in a lower-index matrix, which is not possible in an air-glass structure. However, neither of these two structures has been demonstrated to guide light in a hollow core. The first hollow-core photonic bandgap fibre³² (Fig. 4d) was drawn from a stack of thin-walled silica capillaries with an extra-large hole in the centre, which was formed by omitting seven capillaries from the stack. The strong wavelength dependence of guiding indicated a photonic bandgap effect: only certain frequency ranges fall within the bandgaps and are guided, and so other light quickly leaks out of the hollow core. A similar design has produced the lowest hollow-core loss figure so far¹⁵.

Losses in hollow-core fibres are limited by the same mechanisms that limit loss in conventional fibres: absorption, Rayleigh scattering, confinement loss, bend loss and variations in the fibre structure along the length. However, these losses might be reduced below the levels found in conventional fibres because the majority of the light travels in the hollow core, in which scattering and absorption could be very low. A recent study using a wavelength of 10.6 μm demonstrated that transmission losses could be reduced to well below those of the absorption levels of the bulk materials used in the cladding formation⁹. Reducing the levels to below those of conventional silica fibres remains a challenge. Confinement losses can be eliminated by forming a sufficiently thick cladding. Bend loss is determined by fibre

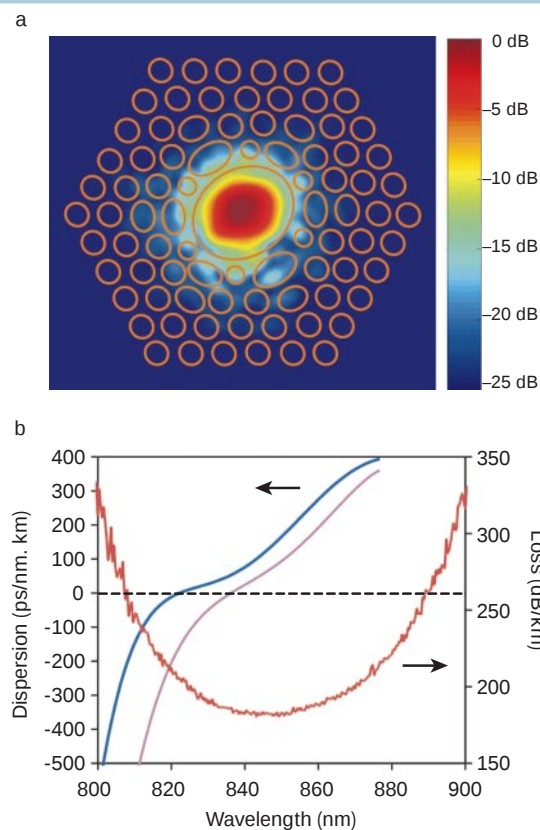


Figure 6 Measured properties of a hollow-core bandgap fibre around 850 nm wavelength. **a**, Logarithmic plot of the near-field pattern at 848 nm after propagation through 60 m of fibre. The locations and shapes of the first few rings of air holes are shown schematically. The light is strongly confined to the core, and the light overlapping with silica is only a tiny fraction of the total. **b**, Measured loss and dispersion characteristics of the fibre. The loss minimum is 180 dB/km at 850 nm wavelength and is probably limited by leakage. The dispersion is low and anomalous over much of the guiding band, which enables soliton propagation. The two lines correspond to the two fundamental polarization modes, which are split because of deformation of the structure.

design and can be reduced to a low level³³ in at least some structures. Rayleigh scattering, like absorption, should be reduced to below the level in bulk fibres, although increased scattering at the many surfaces is a potential problem. The biggest unknown is the level of variation in the fibre structure along the length. Such variations in length scales of less than a micrometre to many metres can be embedded in the middle of the fibre, making them difficult to study directly. It has recently been shown³⁴ that excess loss occurs at wavelengths where coupling of energy occurs from the air-guided mode to confined surface modes. These surface modes have much greater overlap with the glass than the air-guided modes and so experience far greater loss. The dramatic reduction in losses over the past few years suggests that will be reduced still further.

Figure 2c shows an optical micrograph of a hollow-core fibre drawn for use at a wavelength of 850 nm³³. Like most reported hollow bandgap fibres^{15,34,35}, this fibre guides more than a single mode, but other modes become insignificant after a certain propagation length because of poor coupling and higher losses. The near-field pattern of the low-loss mode after 60 m of propagation, excited using a laser source at the input end, is shown logarithmically scaled in Fig. 6a. The figure also shows the locations and approximate shapes of the first few rings of air holes, demonstrating the good confinement of the mode to the core and the low overlap with the glass. The non-circularity of the core and inner air holes is a

result of unintended deformation of the microstructure during the final draw. Light guided in this mode has a minimum loss of 180 dB/km (Fig. 6b, right-hand scale), limited by variations in the fibre microstructure. Coupling to other modes is not observable³³, even when the fibre is wound 100 times around a mandrel with a diameter of 5 mm. Group velocity dispersion in these fibres—a measure of their tendency to lengthen a short pulse during propagation—is anomalous over much of the wavelength band. This implies that these fibres could support short pulse propagation with neither temporal nor spectral distortion as optical solitons. A recent report³⁶ demonstrated soliton-type propagation of 470 nJ pulses over 1.7 m of similar fibre, using 75 fs pulses at a 1,510 nm wavelength. In any conventional fibre, delivery of such pulses is impossible because of the combined effects of dispersion, non-linearity and fibre damage. The Kerr nonlinearity of the hollow-core fibre was low because it was filled with gas, whereas the effects of Raman scattering were eliminated by filling the fibre with xenon. As a result of the balance between the Kerr effect and dispersion, the transmitted pulses suffered virtually no distortion. To optimize this for pulses with a specific pulse length and energy will require tailoring of the dispersion profile of the fibre: there are ample opportunities for such engineering³⁷. These fibres could then deliver ultrashort pulses without distortion over distances of tens or even hundreds of metres in the 800 nm wavelength band as well.

Conclusions and future directions

Hollow-core fibres have massive potential. They might produce lower optical transmission loss than conventional fibres, making them a candidate for future telecommunication transmission systems. They can exhibit substantially higher damage thresholds than conventional fibres⁹, making them suitable for delivery of high-power beams for laser machining and welding. They offer enormous promise for nonlinear optical processes in gases: hollow-core fibre can be construed as a gas cell with diffraction-free and only loss-limited performance. Preliminary experiments in our laboratories have shown that hollow-core fibres filled with hydrogen can act as a cell for stimulated Raman scattering experiments³⁸. We found that the threshold for SRS was orders of magnitude below that in previous experiments. Similar experiments involving ultrashort-pulse propagation in noble gases are currently underway. In a similar spirit, hollow-core fibres could be used for trace gas detection or monitoring, or as gain cells for gas lasers. We have demonstrated delivery of solid particles down a fibre by using optical radiation pressure, and we are developing this for delivery over long distances³⁹. With optical losses likely to fall below 10 dB/km for core diameters of the order of ~10 μ m at optical frequencies, it is anticipated that silica fibre losses in the 2 μ m spectral band will be reduced to levels far below those currently available. In addition, hollow-core fibres from other materials could lead to a new generation of fibres for infrared wavelengths.

The field of PCFs is just coming of age. With optical losses being reduced to—and perhaps beyond—those in conventional fibres and with an understanding of the enormous potential offered by the new structures, fibre optic technology stands ready to be transformed. Future generations might consider ‘conventional’ fibres to be just one example of a low-loss fibre waveguide, and textbooks on fibre optics will draw a far broader picture than they do now. Will optical fibres as we now know them become obsolete? Probably not, but they will become just one string in a far more powerful bow. □

doi:10.1038/nature01940

1. Birks, T. A., Knight, J. C. and Russell, P. St J. Endlessly single-mode photonic crystal fiber. *Opt. Lett.* **22**,

- 961–963 (1997)
2. Birks, T. A., Roberts, P. J., Russell, P. St J., Atkin, D. M. & Shepherd, T. J. Full 2-D photonic bandgaps in silica/air structures. *Elect. Lett.* **31**, 1941–1942 (1995)
3. Yeh, P., Yariv, A. & Marom, E. Theory of Bragg fiber. *J. Opt. Soc. Am.* **68**, 1196–1201 (1978).
4. Knight, J. C., Birks, T. A., Russell, P. St J. & Atkin, D. M. All-silica single-mode optical fiber with photonic crystal cladding. *Opt. Lett.* **21**, 1547–1549 (1996).
5. Kaiser, P. & Astle, H. W. Low loss single material fibers made from pure fused silica. *Bell Syst. Tech. J.* **53**, 1021–1039 (1974).
6. Allan, D. C. et al. Photonic crystal fibers: effective index and band gap guidance. In *Photonic Crystals and Light Localization in the 21st Century* (ed. C. M. Soukoulis) 305–320 (Kluwer, Netherlands, 2001).
7. Monro, T. M. et al. High nonlinearity extruded single-mode holey optical fibers. *FA1*, 1–3 OFC 2002 (2002).
8. Kumar, V. V. R. K. et al. Extruded soft glass photonic crystal fiber for ultrabroad supercontinuum generation. *Opt. Express* **10**, 1520–1525 (2002).
9. Temelkuran, B., Hart, S. D., Benoit, G., Joannopoulos, J. & Fink, Y. Wavelength-scalable hollow optical fibers with large photonic bandgaps for CO₂ laser transmission. *Nature* **420**, 650–653 (2002).
10. Liu, X. et al. Soliton self-frequency shift in a short tapered air-silica microstructure fiber. *Opt. Lett.* **26**, 358–360 (2001).
11. Eggleton, B. J., Kerbage, C., Westbrook, P. S., Windeler, R. S. & Hale, A. Microstructured optical fiber devices. *Opt. Express* **9**, 698–713 (2001).
12. Jasapara, J., Bise, R. & Windeler, R. Chromatic dispersion measurements in a photonic bandgap fibre. *OFC* 2002 519–521 (2002).
13. van Eijkelenborg, M. A. et al. Microstructured polymer optical fiber. *Opt. Express* **9**, 319–317 (2001).
14. Tajima, K., Zhou, J., Nakajima, K. & Sato, K. Ultra low loss and long length photonic crystal fiber. *OFC* 2003 (2003).
15. Venkataram, N. et al. Low loss (13 dB/km) air core photonic band-gap fiber. *PD1.1 Proc. 28th Eur. Conf. Opt. Commun.* (2002).
16. Russell, P. St J. Photonic crystal fibers. *Science* **299**, 385–362 (2003).
17. Knight, J. C. et al. Anomalous dispersion in photonic crystal fiber. *IEEE Photon. Technol. Lett.* **12**, 807–809 (2000).
18. Ferrando, A., Silvestre, E., Miret, J. J. & Andres, P. Nearly zero ultraflattened dispersion in photonic crystal fibers. *Opt. Lett.* **25**, 790–792 (2000).
19. Reeves, W. H., Knight, J. C., Russell, P. St J. & Roberts, P. J. Demonstration of ultra-flattened dispersion in photonic crystal fibers. *Opt. Express* **10**, 609–613 (2002).
20. Ranka, J. K., Windeler, R. S. & Stentz, A. J. Visible continuum generation in air-silica microstructure optical fibers with anomalous dispersion at 800 nm. *Opt. Lett.* **25**, 25–27 (2000).
21. Udem, T., Holzwarth, R. & Hänsch, T. W. Optical frequency metrology. *Nature* **416**, 233–237 (2002).
22. Povazay, B. et al. Submicrometer axial resolution optical coherence tomography. *Opt. Lett.* **27**, 1800–1802 (2002).
23. Ortigosa-Blanch, A. et al. Highly birefringent photonic crystal fibers. *Opt. Lett.* **25**, 1325–1327 (2000).
24. Knight, J. C., Birks, T. A., Cregan, R. F. & Russell, P. St J. Large mode area photonic crystal fibre. *Electron. Lett.* **34**, 1347–1348 (1998).
25. MacPherson, W. N. et al. Remotely addressed optical fibre curvature sensor using multicore photonic crystal fibre. *Opt. Commun.* **193**, 97–104 (2001).
26. Myaing, M. T. et al. Enhanced two-photon biosensing with dual-core photonic crystal fibers. *Opt. Lett.* **28**, 1224–1226 (2003).
27. Limpert, J. et al. High-power air-clad large-mode-area photonic crystal fiber laser. *Opt. Express* **11**, 818–823 (2003).
28. Marcattili, E. A. J. & Schmetzer, R. A. Hollow metallic and dielectric waveguides for long distance optical transmission and lasers. *Bell Syst. Tech. J.* **43**, 1783 (1964).
29. Miyagi, M. & Kawakami, S. Design theory of dielectric-coated circular metallic waveguides for infrared transmission. *J. Lightwave Technol.* **2**, 116–126 (1984).
30. Katagiri, T., Matsuura, Y. & Miyagi, M. Metal-covered photonic bandgap multilayer for infrared hollow waveguides. *Appl. Opt.* **41**, 7603–7606 (2002).
31. Knight, J. C., Broeng, J., Birks, T. A. & Russell, P. St J. Photonic band gap guidance in optical fibers. *Science* **282**, 1476–1478 (1998).
32. Cregan, R. F. et al. Single-mode photonic bandgap guidance of light in air. *Science* **285**, 1537–1539 (1999).
33. Bouwmans, G. et al. Properties of a hollow-core photonic bandgap fiber at 850 nm wavelength. *Opt. Express* **11**, 1613–1620 (2003).
34. Müller, D. et al. Measurement of photonic band-gap fiber transmission from 1.0 to 3.0 μ m and impact of surface mode coupling. *QTL2 Proc. CLEO* 2003 (2003).
35. Johnson, S. G. et al. Low-loss asymptotically single-mode propagation in large-core OmniGuide fibers. *Opt. Express* **9**, 748–779 (2001).
36. Ouzounov, D. G. et al. Generation of high-power, non-frequency shifted solitons in a gas-filled photonic bandgap fiber. *QThPDA3Proc. CLEO* 2003 (2003).
37. Engeness, T. D. et al. Dispersion tailoring and compensation by modal interactions in OmniGuide fibers. *Opt. Express* **11**, 1175–1196 (2003).
38. Benabid, F., Knight, J. C. & Russell, P. St J. Stimulated Raman scattering in hydrogen-filled hollow-core photonic crystal fiber. *Science* **298**, 399–402 (2002).
39. Benabid, F., Knight, J. C. & Russell, P. St J. Particle levitation and guidance in hollow-core photonic crystal fiber. *Opt. Express* **10**, 1195–1203 (2002).

Acknowledgements There have been numerous contributions to my understanding of this topic from my colleagues at Bath, including Philip Russell, Tim Birks, and others. I would like to thank Blazephotonics for help with the fibre illustrated in Figure 2c. This work was partly funded by the UK EPSRC.

Photonic structures in biology

Pete Vukusic and J. Roy Sambles

Thin Film Photonics, School of Physics, Exeter University, Exeter EX4 4QL, UK (e-mail: P.Vukusic@ex.ac.uk)

Millions of years before we began to manipulate the flow of light using synthetic structures, biological systems were using nanometre-scale architectures to produce striking optical effects. An astonishing variety of natural photonic structures exists: a species of Brittlestar uses photonic elements composed of calcite to collect light, *Morpho* butterflies use multiple layers of cuticle and air to produce their striking blue colour and some insects use arrays of elements, known as nipple arrays, to reduce reflectivity in their compound eyes. Natural photonic structures are providing inspiration for technological applications.

The natural world has exploited photonic structures since the Cambrian explosion: the sudden and enormous diversification of life that accompanied the start of the Cambrian period over 500 million years ago. Evidence from this era suggests that the co-development of predator and prey colouration, in step with their visual systems, resulted in the evolution of various life forms. Light can be a significant selection pressure for the evolution of certain animal groups, leading to the astonishing diversity of natural photonic structures present in the world today. Such structures might provide the inspiration for future technological applications.

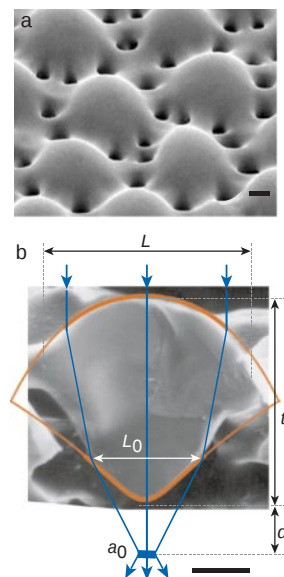
Photonic structures in aquatic systems

Decades before the synthesis of fabricated photonic structures, studies were revealing the complex and elegant way in which it was accomplished naturally. Aquatic systems were the subject of many of the earliest studies. Crystal multilayer structures of guanine were identified as the key component of coloured and broad-band silver reflections in many fish and also found to be a high-reflectance element in many bioluminescent organs and a sub-retinal component that assists with specialized marine vision¹. Recent interest in aquatic systems has revealed photonic structures of astonishing complexity. For example, arm ossicles from light-sensitive species of brittlestar, *Ophiocoma wendtii*, have regular arrays of inorganic microstructures (Fig. 1a) with a characteristic double-lens design² (Fig. 1b). These microlenses are photonic elements, each composed of single anisotropic calcite crystals that focus light towards nerve bundles of photoreceptors 4–7 μm below them inside the arm ossicle tissue. The design of the surface of each lens and the orientation of constituent calcite minimizes spherical aberration and birefringence that might otherwise degrade the optical function. Such microlens arrays assist in creating levels of light and shadow sensitivity that may serve to warn of the presence of predators.

Equally fascinating are the nanostructures found in the hair-like setae of many species of polychaete worm (Fig. 2a). In these creatures, a two-dimensional (2D) hexagonal lattice of voids within the cross-section of each seta (Fig. 2b) creates a natural pseudo-photonic crystal fibre along its full length^{3,4}. The high spatial periodicity of this lattice (Fig. 2c, d) generates a partial photonic bandgap (PBG) by which colour is strongly Bragg-scattered in certain directions. As a consequence of this, strong iridescence is observed laterally. The biological significance of the colour, and therefore of the intricate structure, is understood to be visibility; this contrasts with the principal design purpose of recently

Figure 1 Peripheral layer of ophiocomid brittlestars.

a, Scanning electron micrographs (SEM) of the peripheral layer of a dorsal arm plate from the brittlestar *Ophiocoma wendtii* showing the microlens array. **b**, SEM of an individual lens in *O. wendtii*. The functional region of this lens (L_0) closely matches the profile of a lens that is compensated for spherical aberration (represented by the red lines). The light paths are shown in blue (images reproduced with permission from J. Aizenberg). Bars, 10 μm .



developed synthetic photonic crystal fibres, which is to guide light longitudinally⁵. The absence of centrally located defects within the natural photonic fibres examined so far, prevents light-guiding in these setae.

Photonic structures in insects and birds

Iridescence is much more commonly encountered in terrestrial systems than in aquatic systems. The photonics associated with brightly coloured birds and insects has been extensively studied for over a century but only recently have significant advances been made. Discoveries of partial PBG structures in Coleoptera and Lepidoptera highlight the breadth of nature's innovative use of light. Although certain systems, among them species of Coleoptera⁶ and Hymenoptera⁷, display subtle colouration as a result of diffraction from surface periodicities, the majority of strong photonic effects arise in species that have evolved structures that have layers of alternately high and low refractive index. This leads to optical interference: an effect that pervades much of modern optics and the physics of which is well understood⁸.

In iridescent blue *Morpho rhetenor* butterflies (Fig. 3a), ultralong-range visibility of up to half a mile is attributed to photonic structures formed by discrete multilayers of

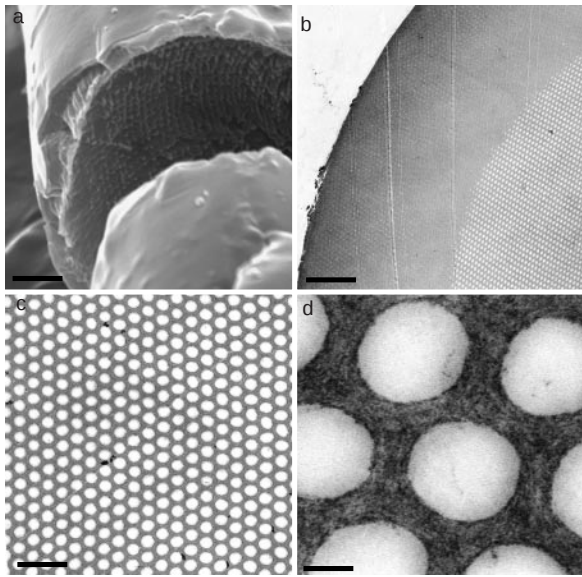


Figure 2 Iridescent setae from polychaete worms. **a**, Scanning electron micrograph (SEM) and **b–d**, transmission electron micrograph (TEM) images of transverse sections through a single iridescent seta. Bars, **a**, 2 μm ; **b**, 5 μm ; **c**, 1 μm ; **d**, 120 nm.

cuticle and air^{9,10} (Fig. 3b, c). Photonic structures of reduced dimensions, present in certain *Colias* butterflies, effect intense UV visibility¹¹. In other species of butterfly, orientational adjustments to the alignment of such discrete multilayers produce strong angle-dependent iridescence that provides high-contrast colour flicker with minimal wing movement¹² or strong iridescence at grazing incidence when viewed posteriorly¹³.

The discrete layering in the examples above contrasts with the more continuous layering, which appears to have developed primarily to induce cryptic colouration, in other butterfly species. In certain architectures, this may not only bring about colour stimulus synthesis¹⁴ but also strong linearly polarized reflection of a specific colour, an effect that contributes to intraspecific communication¹⁵. Several species accomplish this using a multilayered structure embedded in 2D arrays of deep concavities (Fig. 4a, b); this design enables the reflection of yellow light at normal incidence from the base of each concavity and blue light through a double reflection from opposite and perpendicularly inclined sides of each concavity (Fig. 4c) to produce a blue annulus with a yellow centre¹⁶ (Fig. 4d). The juxtaposition of these two colours synthesizes the green colouration perceived by the human eye—and possibly by the predator's.

Certain Coleoptera, however, exhibit continuously layered exocuticle that strongly reflects circularly polarized light through an analogue of optically active cholesteric liquid crystalline structures. The helical arrangement of chitin microfibrils that make up such exocuticle, and which are systematically rotated by a small amount across successive planes, creates a periodicity that produces circularly polarized coloured reflection¹⁷. In other words, the polarized reflection is not derived from optical rotation at a molecular level from the L-amino acids of the cuticle protein and the D-amino sugars of the chitin; instead it arises at the supermolecular level and is similar to that exhibited by a cholesteric liquid crystal from the rotation of the local average alignment direction of the liquid crystal molecules (the director). Although similar helical structures are found in many other iridescent species, they are rarely responsible for similarly strong colouration and anomalous polarization properties⁴.

Structurally coloured avian feather barbs and integument, although they exhibit less structural diversity than scales of Lepidoptera, are no less remarkable. Recent analyses suggest that such

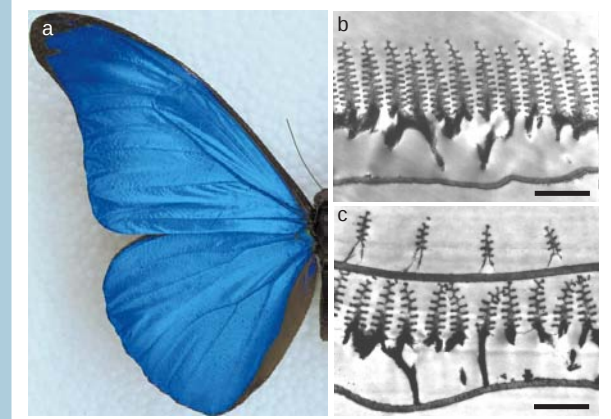


Figure 3 Iridescence in the butterfly *Morpho rhetenor*. **a**, Real colour image of the blue iridescence from a *M. rhetenor* wing. **b**, Transmission electron micrograph (TEM) images showing wing-scale cross-sections of *M. rhetenor*. **c**, TEM images of a wing-scale cross-section of the related species *M. didius* reveal its discretely configured multilayers. The high occupancy and high layer number of *M. rhetenor* in **b** creates an intense reflectivity that contrasts with the more diffusely coloured appearance of *M. didius*, in which an overlying second layer of scales effects strong diffraction⁴. Bars, **a**, 1 cm; **b**, 1.8 μm ; **c**, 1.3 μm .

colour as is seen in many Avian orders, is the product of coherent, rather than incoherent, scatter from the spatial variation in refractive index of medullary keratin in feather barbs or of collagen fibres in the dermis¹⁸.

Photonics in flora

Advanced photonic development is not limited to fauna. Certain anomalous species of flora also show partial PBGs that underpin an often vivid structural colour¹⁹ (Fig. 5a). Invariably this is mediated by variations in 1D multilayering (although more complex structural designs are also thought to exist), producing iridescence in vascular plant leaves, fruits and marine algae⁴. Periodicity is generally formed by laminations of hydrated cellulose, which are usually located close

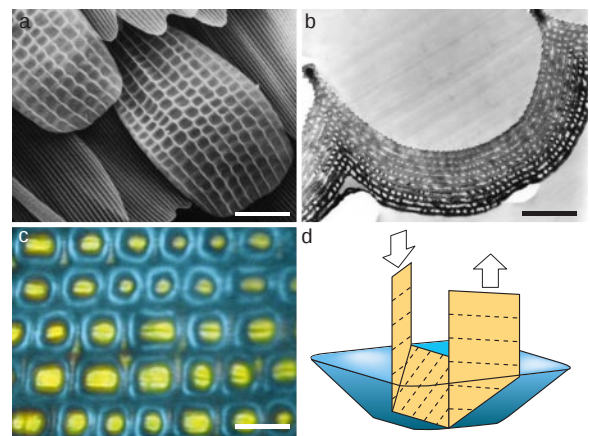


Figure 4 Iridescence in *Papilio palinurus*. **a**, SEM of an iridescent scale showing its array of concavities, each with a section that exhibits the curved multilayering shown by transmission electron microscopy in **b**. This structure produces two simultaneous structural colours **c**, yellow and blue. **d**, The blue annulus is created by a double reflection from opposite and perpendicular concavity sides. **d** also schematically illustrates the way in which incident linearly polarized blue light has its e-vector (dotted lines) rotated by this double reflection. Bars, **a**, 15 μm ; **b**, 1 μm ; **c**, 6 μm .

to epidermal cell walls or within elongated chloroplasts, but it can also be the result of a helical arrangement of cellulose fibrils (Fig. 5b). Iridescence in leaves is thought to produce particular intensity ratios of incident radiation bands that can penetrate to phytochrome centres and thus affect the development of the plant, although the photonic effect may have other functions. Structural colour in fruit skin reduces post-maturation discolouration and ultimately improves dispersal²⁰. Interestingly too, convexly curved epidermal cells on the surface of several particular species of plant appear to act as microlenses²¹, resulting in increased photosynthetic flux density on specially oriented chloroplasts at specific depths within the plants' leaves.

Photonic crystal structures

Natural systems also exist that employ remarkable 3D photonic crystals of cuticle (Fig. 6a, b) to produce partial PBGs, with the effect that bright colour is reflected over a broad angle⁴. From the perspective of modern photonic technology, this indicates an evolutionary step further along the photonic road, since, in principle, 3D periodicity potentially manipulates the flow of light in all directions. The physical structure of such photonic crystals, although varying somewhat between species of examined Lepidoptera, appears consistently to be approximately tetrahedral^{22,23}. Interestingly, bandgap calculations indicate that a perfect tetrahedral configuration offers the highest reflectivity over the broadest angle for a given refractive index contrast between component media²⁴. The physical design of this photonic structure has converged towards one of the most optically efficient configurations. Additionally, the selective advantage of using this microstructure rather than a surface grating or a multilayer system, excluding physiological considerations, is that it can be incorporated into multiple neighbouring domains (Fig. 6c), each orientated slightly differently, to produce a macroscopic colour that is completely independent of the viewing angle⁴.

In contrast to such scattering from a honeycombed lattice of voids described above, other Lepidopteran systems employ randomly orientated solid ellipsoids to produce a multiwavelength scatter that creates the appearance of being white. More eye-catching, however, are the colour centres that arise via Bragg diffraction from 3D lattices of closely packed with solid spheres with high refractive indices. These

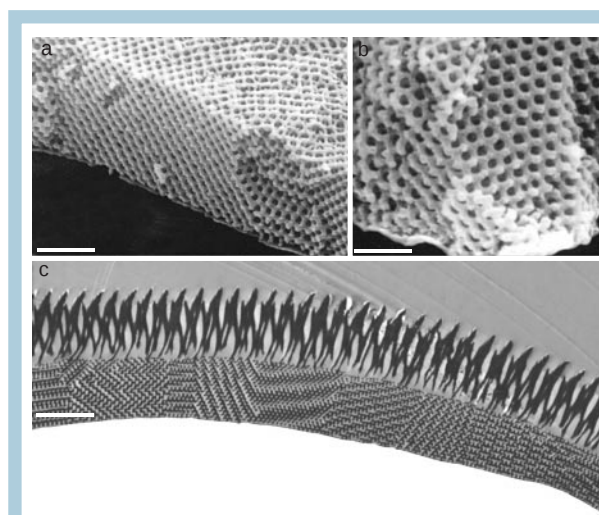


Figure 6 The green colour of *Parides sesostris* is created by a photonic crystal. **a, b**, SEMs of the exposed photonic crystal after the superficial ridging has been removed. **c**, a TEM showing a 50 nm section through the scale shown in **a**; the ridging is intact whereas the neighbouring but differently oriented domains of identical 3D structure are distinguished by contrasting 2D patterns (here, the darkly contrasted material is cuticle). Bars, **a**, 1.2 μm ; **b**, 750 μm ; **c**, 2.5 μm .

lattices have recently been discovered in the iridescent cuticle of certain beetle species (A. R. Parker, V. L. Welch & D. Driver, personal communication), but the effect is far better understood in the inanimate structures associated with gem opals. In these, strong coloured regions are created by aggregates of 150–350 nm diameter amorphous silica spheres arranged hexagonally in layers that are usually stacked randomly. Some opal specimens exhibit parallel domains of ordered packing that are commonly face-centered cubic but which occasionally are hexagonally close-packed²⁵. This ordered structure is understood to be the result of precipitative settling of silica gel solution: the opal's colouring is purely artefactual.

Anti-reflection structures

Although the character of most natural partial PBG systems is associated with bright colour or broad angle reflectivity, a specific form of nanostructure minimizes reflectivity at a surface over broad angles or frequency ranges. The structure gradually matches the optical impedance of one medium with that of its neighbour across the interface. This is invariably achieved by incorporation of arrays of tapered elements, described as nipple arrays, across the boundary. Considerable research into such nipple arrays, which are commonly found on arthropodal ommatidial surfaces (Fig. 7a), suggests a distinct visual selection advantage associated with their use in compound eyes²⁶: the reduced reflectivity from the ommatidial surface enhances the photon collection efficiency of the visual system. The evolutionary development of these anti-reflective nanostructures must have occurred very early—they have been found on the ommatidial surfaces of amber-trapped diptera from the Eocene period²⁷. Further evidence that these photonic structures enhance transmission through interfaces comes from their presence on the dorsal and ventral surfaces of transparent diurnal moth wings²⁸ (Fig. 7b). Additionally, they are seen to remove the Brewster-mediated disparity in reflection of different polarizations from the wing (P.V. and J.R.S., unpublished data), a relevant feature within a forum where visual polarization sensitivity is ubiquitous^{15,29}.

As well as enhancing transparency in certain insects, nanostructures may be associated with the colour black on certain lepidopteran wing scales. Although the dark melanin-based pigmentation in such scales is the absorbing medium, optical scattering

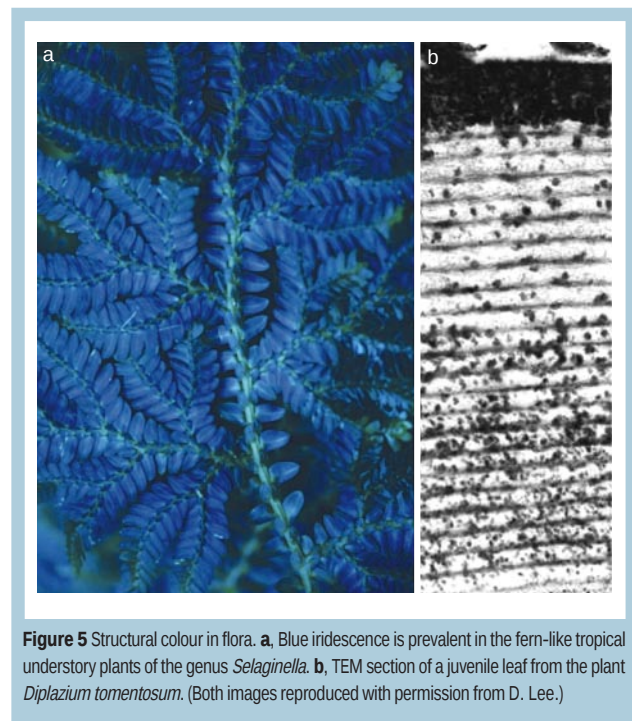


Figure 5 Structural colour in flora. **a**, Blue iridescence is prevalent in the fern-like tropical understory plants of the genus *Selaginella*. **b**, TEM section of a juvenile leaf from the plant *Diplazium tomentosum*. (Both images reproduced with permission from D. Lee.)

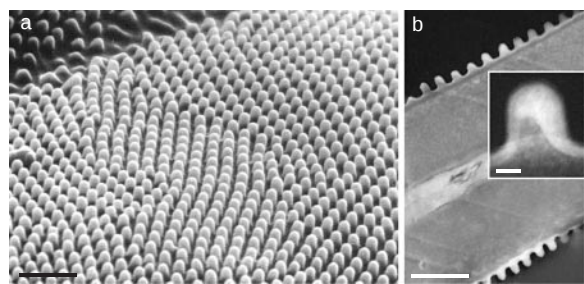


Figure 7 Anti-reflective nipple arrays. **a**, The anti-reflective nipple arrays on ommatidial surfaces of a lepidopteran eye appear identical to those found **(b)** on the transparent wings of certain hawkmoths. This image shows a transparent wing section with a nipple array on both surfaces. The inset image shows the magnified profile of a single nipple. Bars, **a**, 1 μm ; **b**, 1 μm ; inset, 100 nm.

from specific tapered nanostructures on and within each scale significantly increases incident light absorption by the pigment and subsequently enhances the visual appearance of the blackness (P.V., J.R.S. and C. R. Lawrence, unpublished data).

Future considerations

Numerous studies, many of them very recent, have sought to discover and characterize the photonics associated with natural specimens. Many of the revealed designs have existed naturally for millennia, but were, until recently, considered to be the product of contemporary technological innovation. Despite considerable advances in our understanding, significant weaknesses persist. Not enough is known about the material used in these natural systems, particularly with reference to their real and imaginary refractive index component dispersion properties across relevant frequency ranges. Additionally, the sheer physical complexity of many natural systems often renders an accurate representation of their structure extremely difficult. Rigorous application of electromagnetic grating theory³⁰, which increasingly uses modelling in finite element form, is beginning to address some of this complexity.

Natural systems offer technologically unrealized photonic structures and design protocols. Although they have evolved for specific biological purposes and comprise physiologically constrained constituent materials, they provide inspiration for technological applications that require UV, visible, IR or microwave frequency bands. Adaptive replication of certain natural structures, using commercially available materials with better suited optical properties, is already yielding marketable benefit. The range of these replicated structures, and potential future applications, varies, from those associated with advanced optical properties of paints, fabrics and displays to anti-reflection of UV and visible light in electro-optical components and of microwaves from large surfaces, to the intrinsic photonic lasing and light-guiding components associated with optical processing. Present manufacturing techniques such as thin layer vacuum deposition, template embossing and various forms of lithography are limited and often prohibitively expensive. It is for this reason that the

biological processes of manufacture are of great interest. Invariably, these structures are self-assembled in a range of environments—in pupae, eggs or during early growth—through design templates that are often influenced by environmental factors³¹ but are ultimately controlled at the genetic level. Although many natural photonic structures may not be technologically appropriate for direct use, genetic manipulation may one day yield directly exploitable photonic components. □

doi:10.1038/nature01941

- Denton, E. J. Reflectors in fishes. *Sci. Am.* **224**, 64–72 (1971).
- Aizenberg, J., Tkachenko, A., Weiner, S., Addadi, L. & Hendler, G. Calcitic microlenses as part of the photoreceptor system in brittlestars. *Nature* **412**, 819–822 (2001).
- Parker, A. R., McPhedran, R. C., McKenzie, D. R., Botten, L. C. & Nicorovici, N. A. Photonic engineering: Aphrodite's iridescence. *Nature* **409**, 36–37 (2001).
- Vukusic, P. in *Optical Interference Coatings* (eds Kaiser, N. & Pulker, H. K.) 1–34 (Springer, New York, 2003).
- Knight, J. C. & Russell, P. St. J. Photonic Crystal Fibers: New Ways to Guide Light. *Science* **296**, 276–277 (2002).
- Parker, A. R. Diffracting optics in animals: diversity and biological significance. *Eur. Opt. Soc. Top. Meet. Digest Series* **18**, 27–30 (1998).
- Hinton, H. E., Gibbs, D. F. & Silberglied, R. Stridulatory files as diffraction gratings in mutillid wasps. *J. Insect Physiol.* **15**, 549–552 (1969).
- Kaiser, N. & Pulker, H. K. (eds) *Optical Interference Coatings*. (Springer, New York, 2003).
- Vukusic, P., Sambles, J. R., Lawrence, C. R. & Wootton, R. J. Quantified interference and diffraction in single *Morpho* butterfly scales. *Proc. R. Soc. Lond. B* **266**, 1402–1411 (1999).
- Kinoshita, S., Yoshioka, S. & Kawagoe, K. Mechanisms of structural colour in the *Morpho* butterfly: cooperation of regularity and irregularity in an iridescent scale. *Proc. R. Soc. Lond. B* **269**, 1417–1421 (2002).
- Ghiradella, H. Light and colour on the wing: Structural colours in butterflies and moths. *Appl. Opt.* **30**, 3492–3500 (1991).
- Vukusic, P., Sambles, J. R., Lawrence, C. R. & Wootton, R. J. Structural colour: Now you see it – now you don't. *Nature* **410**, 36 (2001).
- Lawrence, C. R., Vukusic, P. & Sambles, J. R. Grazing incidence iridescence from a butterfly wing. *Appl. Opt.* **41**, 437–441 (2002).
- Agoston, G. A. *Colour Theory and its Applications in Art and Design* (Springer, New York 1987).
- Sweeney, A., Jiggins, C. & Johnsen, S. Polarized light as a butterfly mating signal. *Nature* **423**, 31–32 (2003).
- Vukusic, P., Sambles, J. R. & Lawrence, C. R. Colour mixing in wing scales of a butterfly. *Nature* **404**, 457 (2000).
- Neville, A. C. & Caveney, S. Scarabeid beetle exocuticle as an optical analogue of cholesteric liquid crystals. *Biol. Rev.* **44**, 531–562 (1969).
- Prum, R. O. & Torres, R. H. Structural colouration of avian skin: convergent evolution of coherently scattering dermal collagen arrays. *J. Exp. Biol.* **206**, 2409–2429 (2003).
- Lee, D. W. Iridescent blue plants. *Am. Sci.* **85**, 56–63 (1997).
- Lee, D. W. Ultrastructural basis and function of iridescent blue color of fruits in *Elaeocarpus*. *Nature* **349**, 260–262 (1991).
- Bone, R. A., Lee, D. W. & Norman, J. M. Epidermal cells functioning as lenses in leaves of tropical rain-forest shade plants. *Appl. Opt.* **24**, 1408–1412 (1985).
- Vukusic, P. Shedding light on butterfly wings. *Proc. SPIE* **4438**, 85–95 (2001).
- Argyros, A., Manos, S., Large, M. C. J., McKenzie, D. R., Cox, G. C. & Dwarde, D. M. Electron tomography and computer visualisation of a three-dimensional 'photonic' crystal in a butterfly wing-scale. *Micron* **33**, 483–487 (2002).
- Chan, C. T., Ho, K. M. & Soukoulis, C. M. Photonic band gaps in experimentally realizable periodic dielectric structures. *Europhys. Lett.* **16**, 563–568 (1991).
- Sanders, J. V. & Darragh, P. J. The microstructure of precious opal. *The Mineralogical Record* **2**, 261–268 (1971).
- Land M. F. & Nilsson, D. E. *Animal Eyes*. (Oxford Univ., Oxford, 2001).
- Parker, A. R., Hegedus, Z. & Watts, R. A. Solar-absorber antireflector on the eye of an Eocene fly (45 Ma). *Proc. Roy. Soc. B* **265**, 811–815 (1998).
- Yoshida, A., Motoyama, M., Kosaku, A. & Miyamoto K. Antireflective nanoprotuberance array in the transparent wing of a hawkmoth *Cephanodes hylas*. *Zool. Sci.* **14**, 737–741 (1997).
- Wehner, R. Polarization vision – a uniform sensory capacity? *J. Exp. Biol.* **204**, 2589–2596 (2001).
- Gralak, B., Tayeb, G. & Enoch, S. *Morpho* butterflies wings color modeled with lamellar grating theory. *Opt. Express* **9**, 567–578 (2001).
- McMillan, W. O., Monteiro, A. & Kapan, D. D. Development and evolution on the wing. *Trends Ecol. Evol.* **117**, 125–133 (2002).

Contacts
Publisher: Ben Crowe
Editor: Paul Smaglik
Marketing Manager: David Bowen

**European Head Office,
London**

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Global Head of Naturejobs:
Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)
Andy Douglas (4975)
Frank Phelan (4944)

**Netherlands/ Italy/
Iberia/ Belgium:**

Evelina Rubio Hakansson (4973)
Scandinavia: Sille Opstrup (4994)

France/ Switzerland:
Amelie Pequignot (4974)

Natureevents:
Paul Constant (4954)

Production Manager: Billie Franklin

To send materials use London
address above.
Tel +44 (0) 20 7843 4814
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

International

Advertising Coordinator:
Hind Berrada (4935)

Naturejobs web development:
Tom Hancock

Naturejobs online production:
Ben Lund

European Satellite Office

Germany/ Austria:

Patrick Phelan, Odo Wulffen
Tel + 49 89 54 90 57 11/-2
Fax + 49 89 54 90 57 20
e-mail: p.phelan@nature.com
o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707
Tel +1 800 989 7718
Fax +1 800 989 7103
e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

US Advertising Coordinator:
Linda Adam

Japan Head Office, Tokyo

MG Ichigaya Building (5F),
19-1 Haraikatamachi,
Shinjuku-ku,
Tokyo 162-0841
Tel +81 3 3267 8751
Fax +81 3 3267 8746

Asia-Pacific Sales Director:
Rinoko Asami
e-mail: r.asami@naturejpn.com

naturejobs

Upping the ante

A bold move by the Russell Group, which encompasses 19 of Britain's largest universities, could make the country a more attractive place for scientists. The group, which includes Cambridge, Oxford and Imperial College London, last month called for an increase in the salary scale for high-level appointments at its members' institutions. Such a move, if enacted, could halt — or even reverse — Britain's brain drain by offering competitive salaries to elite scientists.

The group's plan is not without precedent. Science Foundation Ireland's infusion of funds into the country for both research infrastructure and grants, for example, has shown that higher salaries can help to lure the elite to your shores.

But the Russell Group's vision is contentious because it would mean that students at those institutions would have to pay higher fees. It could also cause resentment from professors at colleges not in the group who would still be receiving the lower national salary scale. The plan may be successful if it drives up the national salary scale or at least makes it more flexible — few would argue with the idea that top scientists should be paid their market value. But it could fail if it spreads the wealth disproportionately to a few high-fliers, while leaving the bulk — both within and outside the Russell Group — with lower pay.

The problem, as always, is finding the money. Who will finance the market-value salaries? Putting the burden on students — who are already strained by taking on student loans — seems unfair. But the government surely doesn't have an unlimited amount of cash for university salaries. Perhaps a system of private endowments, which have buoyed private US universities such as Harvard and Stanford, could help. Although professors may find that the market value for their skills could be just as volatile as the price of stocks.

Paul Smaglik
Naturejobs editor



Contents

SPECIAL REPORT

Building a better
scientific rapport

p858

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

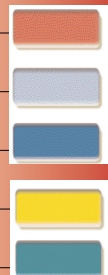
FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS



Bold architecture aims to give science a posh home, spur interdisciplinary rendezvous and reel in big names. But does it matter to those who work there? Kendall Powell finds out.

Charlie Hart, a protein chemist and head of investor relations at ZymoGenetics in Seattle, looks out of his window and describes the scene: “A 35-foot cruiser is going by and a float plane is getting ready to land.” He says that the view from one of the most unusual research facilities in the United States — a renovated steam power plant, *circa* 1914 — is one reason why the 365-person company had only a 4% staff turnover last year.

Other factors include large sunlit laboratory spaces, decks built for lunching and receptions, exhibitions of local artists’ work and a public dock on Lake Union where employees can launch kayaks, go windsurfing or crew boats.

Planners at other companies and research institutes have also come to recognize that amenities and building design can work together to give workers a top-quality environment for doing science. And recruiters don’t shy away from touting these atmospheric perks to prospective new scientists.

Architecture and facilities can play a critical role in drawing people to new endeavours. They can also drive scientific communication and collaboration to new interdisciplinary levels (see page 718).

MOVE WHERE?

A good work environment and research tools can be especially important when trying to lure scientists to new institutions in places outside well-established scientific corridors such as Boston and San Francisco in the United States or Cambridge in Britain.

The Stowers Institute for Medical Research in Kansas City, Missouri, sets the tone for visiting scientists’ first impressions, with facilities accented in exotic woods, limestone and slate. Chief executive Bill Neaves says that the founders — two cancer survivors who pledged more than \$1 billion for basic research in their home town — wanted to build the finest research environment possible. The 10-acre campus will eventually house 40 investigators.

Apart from having core facilities in bioinformatics, proteomics, imaging and transgenics staffed by rising stars, the institute also includes 13 extravagant onsite hotel suites for guests, a robotically maintained mouse house, and liquid nitrogen piped directly into each lab. Although some may scoff at the expense of what many scientists would consider to be frivolous details,



MAX PLANCK INSTITUTE

Kai Simons feels that good architecture can attract scientists.

Neaves sees the ambience as critical for recruitment.

“The bottom line for folks is whether they are convinced they can do their best work here,” Neaves says. “Your success will not be limited by the quantity or quality of lab facilities.” He points out that of the 18 faculty members already on board, three were recently elected into the American Academy of Arts and Sciences — a first for Kansas City — and two resigned from their Howard Hughes Medical Institute (HHMI) posts to move. It doesn’t hurt that the Stowers pays more than the HHMI. Neaves was also able to recruit one scientist from as far afield as France.

Not that Europe is lacking innovative design and its associated benefits. In Germany, the Max Planck Institute for Molecular Cell Biology and Genetics in Dresden is housed in a building recognized for its architectural boldness. This has played a role in bringing scientists to an area not known for its science. Kai Simons, its executive director, says that he was surprised by how fast positions and graduate-programme spots filled up.

“We had to make our case because we were moving into a scene that had to be built and reshaped. The building makes an impact on people,” says Simons. There are fewer tenure-track positions in east Germany compared with the west, but Simons notes that the institute is helping to redress this balance by



Bold move: applicants are flocking to the innovative Max Planck Institute for Molecular Cell Biology and Genetics in Dresden.

providing positions for young scientists in the area.

In Vienna, the Institute of Molecular Biotechnology (IMBA) hopes to use architecture to attract the well-educated scientists of central and eastern Europe. Next door to the high-powered Research Institute of Molecular Pathology (IMP), the IMBA's director Josef Penninger would like to copy his neighbour's success. "No one wants to leave the IMP, which is the best indicator of a good research institute," he says.

Penninger jumped at the chance to work on the IMBA's floorplans and design — an opportunity that helped recruit him from the Amgen Institute in Canada. "How often do you get the chance to create your own research institute from scratch?" he asks. After inspecting the plans, he added a bottom floor for a mouse facility and greenhouses on the top floor. He requested bamboo trees to flank the lab balconies in the central atrium, too.

TRADING SPACES FOR FACES

But can expensive materials, famous architects and posh furniture alone attract world-class scientists? It's an open question, says Susan Marqusee, a structural biologist at the University of California, Berkeley. Marqusee is also associate director of the California Institute for Quantitative Biomedical Research (QB3), a joint programme split between the University of California campuses at San Francisco, Berkeley and Santa Cruz. She says that the first priority for QB3 researchers is to have more space, but a place that fosters communication among staff is a close second.

Robert McGhee, laboratory architect for HHMI, says, "The number one thing for people has nothing to do with facilities, but rather who's on campus, who can

they collaborate with. You can have the finest facilities, but if you don't have recruitment, it doesn't matter." He is working with architect Rafael Viñoly on HHMI's new Janelia Farm campus, to be built into the sloping hills in Ashburn, Virginia.

In a survey of his colleagues, Penninger says he found that scientists adapt to whatever environment they find themselves in and are fairly indifferent to their physical surroundings. But human-resource directors say that a high-quality atmosphere, combining cutting-edge core facilities with a pleasant workplace, exerts a magnetic force.

Planners of the AstraZeneca R&D facility in Waltham, Massachusetts, counted on atmosphere and workplace comforts to recruit employees. The cluster of five buildings set in a star shape is nestled among trees and wetlands in the Boston suburbs.

"It's really a cool place, absolutely stunning right when you walk in," says Peter Laskey, the company's human-resources director. The goal of the design and location was to have researchers in a natural setting, close to affordable housing and keeping close ties with Boston's scientific community.

The attractiveness and natural setting of a site, along with open, light-filled laboratories and comfortable meeting spaces, can allow scientists to get on with their work and still be well connected with their colleagues and the outside world. In Seattle, Charlie Hart says that watching the lake activity outside the window while chatting on the phone is a major reason he and many others have been with ZymoGenetics for 20 years and counting.

Kendall Powell is a freelance science writer based in Broomfield, Colorado.